INCORPORATING RISK AND UNCERTAINTY INTO FORECASTS OF WATERBORNE TRAFFIC FLOWS

A REFERENCE MANUAL OF METHODOLOGIES AND HYPOTHETICAL EXAMPLES

June 1997

INCORPORATING RISK AND UNCERTAINTY INTO FORECASTS OF WATERBORNE TRAFFIC FLOWS

A REFERENCE MANUAL OF METHODOLOGIES AND HYPOTHETICAL EXAMPLES

by:

Jack C. Kiefer

of:

Planning and Management Consultants, Ltd.
6352 South U.S. Highway 51
P.O. Box 1316
Carbondale, Illinois
(618) 549-2832

A Report Submitted to:

Institute for Water Resources
Casey Building
7701 Telegraph Road
Alexandria, Virginia

June 1997

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This report was completed by Planning and Management Consultants, Ltd. (PMCL) under the Planning Methodologies research work unit at the Corps of Engineers (COE), Institute for Water Resources (IWR). The Planning Methodologies research program, is under the direction of Mr. Michael R. Krouse, Chief of Technical Analysis and Research Division at IWR. Mr. Keith Hofseth served as principal investigator for the research unit for this project.

The author deeply appreciates the professional guidance of Mr. Keith Hofseth and Dr. David Moser of the Institute of Water Resources. Their review comments and feedback greatly contributed to the quality of this manual. Thanks are also extended to Ms. Mona King of IWR for her helpful suggestions on the final report.

Finally, the author would like to thank Dr. Jerzy Kocik for his meaningful comments and input to the report and Ms. Stephennie Brumley and Ms. Lisa Gallatin for their patient and skillful preparation of this text. As always, any deficiencies in the report remain the responsibility of the authors.

I. INTRODUCTION

BACKGROUND

The forecasting of river borne commodity flows is essential to analyses of the economic benefits of the Nation's inland navigation system. Typically, forecasts of commodity flows rely on the development of point estimates of the amount of tonnage or number of barge tows that will pass a particular lock or system of locks over time. These point estimates represent expectations, whether they be statistical expectations (e.g., averages) or expectations based on professional judgement. Reference to fundamental principals of probability, however, indicate that the likelihood these point forecasts will actually come true is zero. In order to qualify expectations of the future, then, it is necessary to develop forecast intervals within which actual future flows will be expected at a specified level of confidence. This is where risk and uncertainty analysis comes into play.

Put simply, risk is the probability of suffering economic and other types of loss. Uncertainty, may be more broadly defined as the probability of being incorrect. It follows, then, that situations that involve risk are a subset of situations that involve uncertainty. Situations of uncertainty are translated into situations of risk when one assigns consequences (i.e., costs) to being incorrect. The concept of confidence is inversely related to risk and uncertainty. High levels of confidence in a decision or an outcome of a decision correspond to lower levels of risk and uncertainty, and vice versa.

The Principles and Guidelines (Water Resources Council, 1983) identifies the need to examine and determine levels of risk and uncertainty. According to the Principles and Guidelines (P&G), the planner's primary role in dealing with risk and uncertainty is:

> *to identify the areas of sensitivity and describe them clearly so that decisions can be made with knowledge of the degree of reliability of available information.*

P&G provides limited guidance on how to measure and portray risk and uncertainty in any particular context (e.g., in forecasting commodity flows). P&G recommends the use of objective and subjective probability distributions where possible, and advises at a minimum the use of sensitivity analysis, which tests the sensitivity of outcomes with respect to variation in the magnitude of key parameters or assumptions.

PURPOSE AND OBJECTIVES

In support of inland navigation analysis, this study aims at developing a manual that will help planners incorporate risk and uncertainty analysis into forecasts of commodity flows. The objective of this manual is to incorporate risk and uncertainty analysis into four basic methodologies that the

Corps historically has used to forecast commodity flows. According to *A Review of 16 Planning and Forecast Methodologies* (Grier and Skaggs, 1992), these methodologies may be summarized as:

(1) The application of independently derived commodity-specific annual growth rates to base traffic levels.

(2) Shipper surveys of existing and potential waterway users to determine future plans to ship by barge.

(3) Statistical analysis using regression and correlation analysis to predict future waterborne traffic based on independent economic variables.

(4) Detailed long-range commodity supply-demand and modal split analysis incorporating the production and consumption patterns of individual economic regions within the waterway hinterland.

An example of preparing a forecast that incorporates risk and uncertainty analysis is provided for each of these methodologies. Each of the examples are formulated for a hypothetical river segment that has a single lock and dam facility. The hypothetical examples incorporate agricultural commodities that are common to the inland waterway system.

This manual is not a guidebook for forecasting or a statistics text. Readers are referred to the following texts for help in absorbing the material that is presented. This is certainly not an exhaustive list, but be assured that all are good:

- Econometric Models & Economic Forecasts (Pindyck and Rubinfeld, 1981)
- Elements of Econometrics (Kmenta, 1986)
- Statistical Analysis for Business and Economics (Harnett and Murphy, 1985)
- A Guide to Econometrics (Kennedy, 1992)

The users of this manual should also become familiar with statistical software packages that facilitate analyses of risk and uncertainty. The examples that are provided herein were developed to a large extent through the combined use of SAS©, @Risk©, and BestFit© software. These packages also have good documentation that describes the statistics behind the output. Information on who to contact for the lease or purchase of these packages can be found in the references section of the manual.

HOW TO USE THIS MANUAL

Because these procedures are applied to a hypothetical scenario, the reader may find that the provided forecasting examples are too simplistic or too complex for his or her own particular forecasting requirements. It must be kept in mind, however, that the purpose of this manual is not to describe how to develop a forecast. Rather, this manual should be used as a reference to identify sources of risk and uncertainty that are common to most Corps commodity forecasting exercises. The manual also should be used as a resource for identifying available procedures that may be used to quantify risk and uncertainty. In essence, this manual should serve to illustrate the types of analyses that should be undertaken when preparing forecasts of waterborne traffic, whether the forecasts concern movements of agricultural commodities, nonagricultural commodities, or both.

ORGANIZATION OF MANUAL

The next chapter describes the hypothetical scenario to which the four basic forecasting methodologies are applied. The discussion introduces the Oak River and Chadwick Lock and Dam, and outlines the study area, including a principal destination for waterborne traffic and major shipping origins. Major assumptions about the agricultural base of the hypothetical study area are discussed.

Table I-1 below summarizes the forecasting methods that are reviewed in this manual along with the methods that are used to improve the forecasts by incorporating risk and uncertainty.

Chapter III introduces risk and uncertainty analysis into the first of four forecasting methods, namely, application of commodity-specific growth rates to base traffic levels. Some useful mathematical and statistical rules and assumptions are used first to derive forecast intervals for individual commodities, and then to combine the individual forecasts into a forecast of total grain shipments.

TABLE I-1

SECTIONS OF MANUAL

Chapter	Forecasting Method	Simple Description	Sources of Uncertainty/Error	Method of Improvement
III	Growth Rates	Growth rates applied to base traffic levels of specific commodities	Assumed growth rates and base traffic levels	Use of historical variation to develop probability distribution of forecast values
IV	Shippers Survey	Survey of shippers with regard to plans to ship by barge	Errors in shippers expectations	Subjective probability exercise and use of Normal distribution
V	Regression Analysis	Estimation of numeric relationship to explain changes in traffic levels	Random, sampling, conditioning, and specification error	Construction of statistical confidence intervals together with Monte Carlo simulation
VI	Top-Down Approach	Identification and quantification of all factors considered to affect traffic levels	Errors in assumptions and random, sampling, conditioning, and specification error	Simulation of system of assigned probability distributions and statistical regression models

In Chapter IV, a sample shipper survey instrument is developed. The survey is designed to elicit from a group of hypothetical shippers plans for shipping by barge, based on discrete subjective probability levels. Through the use of the mathematical techniques introduced in Chapter III, the survey results of individual shippers are used to develop estimates of expected total grain shipments, as well as estimates of the variation around this total.

Chapter V forecasts barge movements by the Chadwick Lock using regression analysis. The concept of statistical confidence is discussed, and confidence intervals for predicted tow movements are developed using standard regression procedures. The concepts of random, sampling, and specification error are also introduced and explained.

Chapter VI formulates an example of forecasting waterborne traffic using a detailed supply and demand analysis which is also known as the top-down approach. The incorporation of risk and uncertainty analysis into the top-down approach is undertaken with the help of Monte Carlo simulation. Ways to cope with and quantify risk and uncertainty are described at each level of the top-down analysis, from the amount of grain harvested to the amount of grain passing the Chadwick Lock.

Chapter VII presents a discussion on the choice of forecasting methodology and on the constraints and potential tradeoffs that exist in forecasting waterborne traffic levels. The chapter concludes with some basic rules to follow when confronting the task of forecasting with risk and uncertainty.

II. THE HYPOTHETICAL STUDY AREA

Figure II-1 illustrates the hypothetical study area that serves as the geographical setting for the application of the four forecasting methodologies. The Oak River is the main waterway in the region and stretches 600 miles from its headwaters north of Evanstown to its confluence with Anderson Bay at Cajun City. The Oak River has two tributaries, The Pitt Agricultural Canal and the Little Oak River. The Oak River and its tributaries support barge traffic year-round under most weather conditions. The river has been closed to barge traffic only one time in the last 100 years—one week during the severe drought of 1988. The Oak River Basin has a predominantly agricultural economic base. In terms of tonnage, grains are by far the most widely transported commodity group on the Oak River and its tributaries. Therefore, shipments of grains are the focus of the waterborne traffic forecasts of the following chapters.

This manual focuses on the risk and uncertainty involved in forecasting barge traffic passing the Chadwick Lock, the only lock and dam facility on the studied reach of the Oak River. The agricultural production of the Evanstown business economic area (BEA) supports the use of the Oak River and the Chadwick Lock. The largest cities in the Evanstown BEA are Evanstown, Jackson, and Franklin. These cities serve as the primary shipping points for grains on there way to foreign export. These cities are also home to large grain processing facilities. Large plants and distilleries in Evanstown process raw grain into food and beverage products. In Franklin, facilities convert raw grain into ethanol. Meanwhile, large dairy operations in Jackson demand raw grain for feed. Infra-regional shipments of grain are typically transported by truck or rail to the shipping and processing facilities in Evanstown, Jackson, and Franklin.

The primary export facility and destination for exported grains is located outside the Evanstown BEA at Cajun City. All grain shipments that pass the Chadwick Lock are assumed to be en route to Cajun City. Grains may also be shipped to Cajun City via the Gulf Central and Atlantic Coast Railroads, which creates a modal choice for transport of grains.

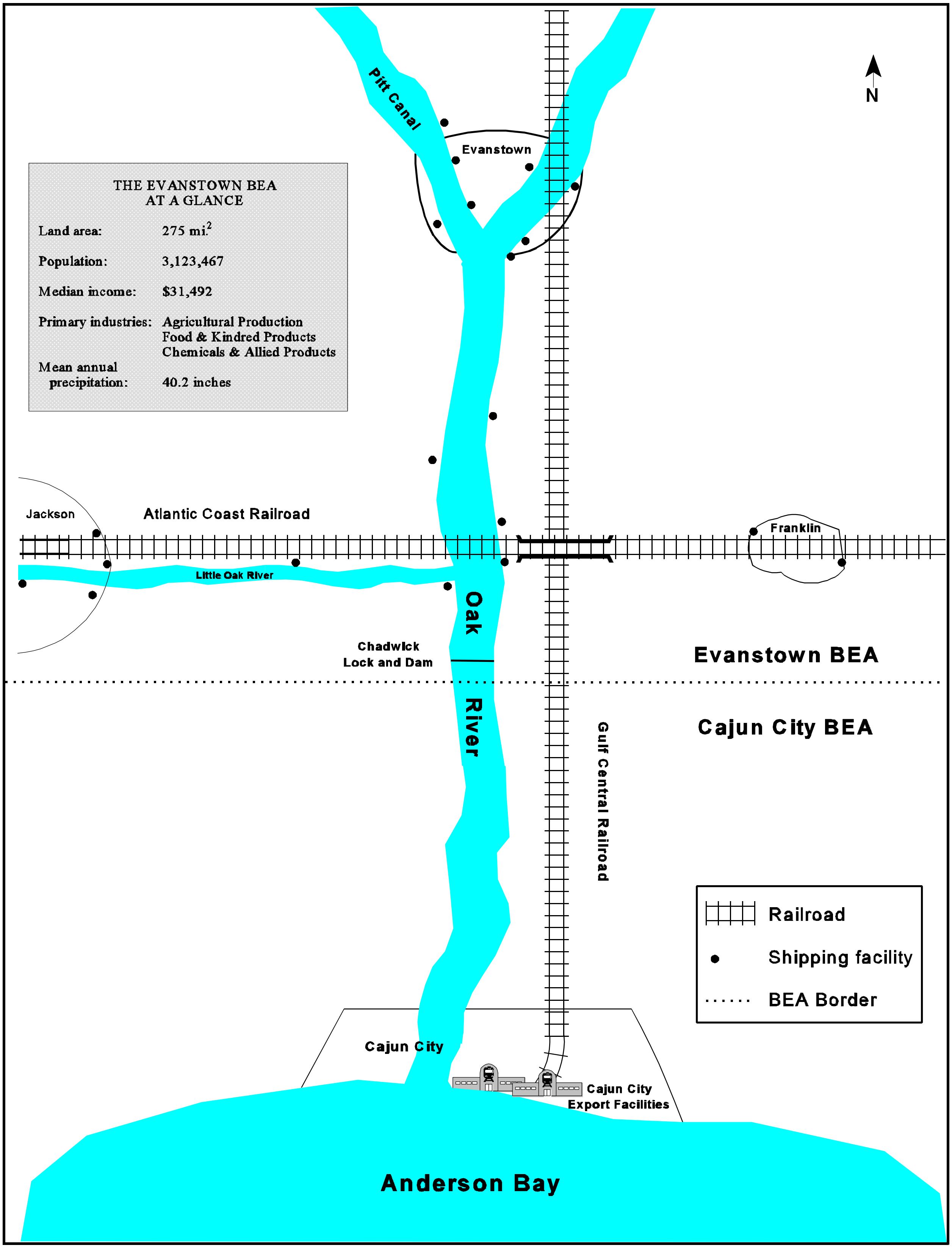

FIGURE II-1
MAP OF THE HYPOTHETICAL STUDY AREA

III. FORECASTING USING GROWTH RATES

One method of forecasting commodity flows entails the application of independently derived commodity-specific (e.g., corn, wheat, soybeans, and other grains) growth rates to given starting (or base) shipment levels to forecast future commodity flows. The amount or volume of a commodity passing a lock is based on an estimate of the base volume of shipments multiplied by commodity-specific growth rates. That is, for a particular forecast year and commodity, the commodity forecast is determined from:

$$VS_{f,i} = B_i \quad (1 + G_i)^f \tag{3.1}$$

where

VS	=	volume shipped by use of waterborne transport (thousands of tons)
B	=	base volume of ith commodity shipped (thousands of tons)
G	=	commodity-specific growth rate (decimal fraction)
f	=	subscript and superscript denoting any future year (f = 0,1,2, ...n, where f=0 denotes the current or starting period with base shipments B and n denotes the number of periods to the forecast horizon)
i	=	subscript denoting the specific commodity

The value of B in Equation 3.1 is usually taken to represent the volume shipped in a recent year or an average of shipments over a few recent years. The value of G is typically fixed at a conservative long-run rate of growth. Because B and G are assumed to retain specific values, the forecast of commodity flows (VS) is said to be deterministic, meaning there are no random (or *stochastic*) elements to consider. At least implicitly, the estimates of B_i and G_i are assumed to be known with certainty.

EXAMPLE APPLICATION

The application of this approach to forecasting commodity flows requires assumptions for the parameters B and G for each commodity under consideration. To illustrate the application of this approach, consider the data reported in Table III-1 for historical annual grain shipments moving past the hypothetical Chadwick Lock. The table reports annual tonnages by crop, the average of shipments over the last ten years, and the average annual rate of growth in shipments over the historical period. This is enough information to prepare a forecast using this simple method.

Assume for now that the base volume selected for the commodities is the average annual amount shipped passed Chadwick Lock over the 1986-1995 period. For example, for corn, the parameter B in Equation (3.1) is set to a value of 15,512. Also assume that the average annual

TABLE III-1 HISTORICAL GRAIN SHIPMENTS CHADWICK LOCK (1970-1995) (THOUSANDS OF TONS)					
Year	**Corn**	**Wheat**	**Soybean**	**Other Grain**	**Total**
1970	12,029	350	1,137	159	13,674
1971	13,304	657	1,181	357	15,498
1972	10,557	512	839	124	12,032
1973	12,257	677	1,312	255	14,502
1974	11,037	873	1,304	278	13,492
1975	14,631	1,066	1,331	179	17,208
1976	11,998	1,139	979	196	14,312
1977	13,225	1,190	1,551	174	16,140
1978	17,427	802	2,062	175	20,466
1979	19,289	925	2,190	191	22,596
1980	16,821	1,075	1,800	251	19,947
1981	21,219	1,620	1,727	163	24,729
1982	18,410	1,152	1,673	208	21,442
1983	7,892	833	1,543	214	10,482
1984	15,042	1,123	1,698	246	18,110
1985	21,850	1,317	2,501	252	25,920
1986	16,732	872	2,309	180	20,094
1987	12,581	860	2,117	229	15,787
1988	8,462	740	1,604	96	10,901
1989	17,341	1,516	2,456	222	21,535
1990	17,451	1,563	2,007	181	21,202
1991	15,145	776	2,350	104	18,375
1992	19,953	1,452	2,119	164	23,689
1993	12,698	910	1,860	105	15,572
1994	20,772	757	2,534	143	24,205
1995	13,981	951	2,828	103	17,864
Avg. Annual Tonnage (1986-1995)	15,512	1,040	2,218	153	18,922
std. deviation (1986-1995)	3,722	333	353	50	4,094
Avg. Annual Rate of Growth (1970-1995)	0.0062	0.0661	0.0572	-0.0134	0.0118

growth rate in shipments over the 1970-1995 period is selected as the parameter G. For example, for corn, G in Equation (3.1), is set to a value of 0.0062 (=0.62 percent). Referring to equation (3.1) and Table III-1, the application of this approach yields the following series of forecasting equations:

$$\begin{aligned} VS_{corn} &= 15{,}512\ (1 + 0.0062)^{f} \\ VS_{soyb} &= 2{,}218\ (1 + 0.0572)^{f} \\ VS_{wheat} &= 1{,}040\ (1 + 0.0661)^{f} \\ VS_{other} &= 153\ (1 - 0.0134)^{f} \end{aligned} \tag{3.2}$$

where for any particular forecast year, *f*, the total amount of grains passing the Chadwick Lock is estimated as:

$$VS_{total} = VS_{corn} + VS_{soyb} + VS_{wheat} + VS_{other} \tag{3.3}$$

Using the parameters defined above, Table III-2 presents the results of a 20-year forecast of grain shipments (1996 is the year represented by f=0). These results define point estimates of future conditions, but do not incorporate any of the fluctuation that is inherent in the historical data of Table III-1. The forecast is illustrated in Figure III-1.

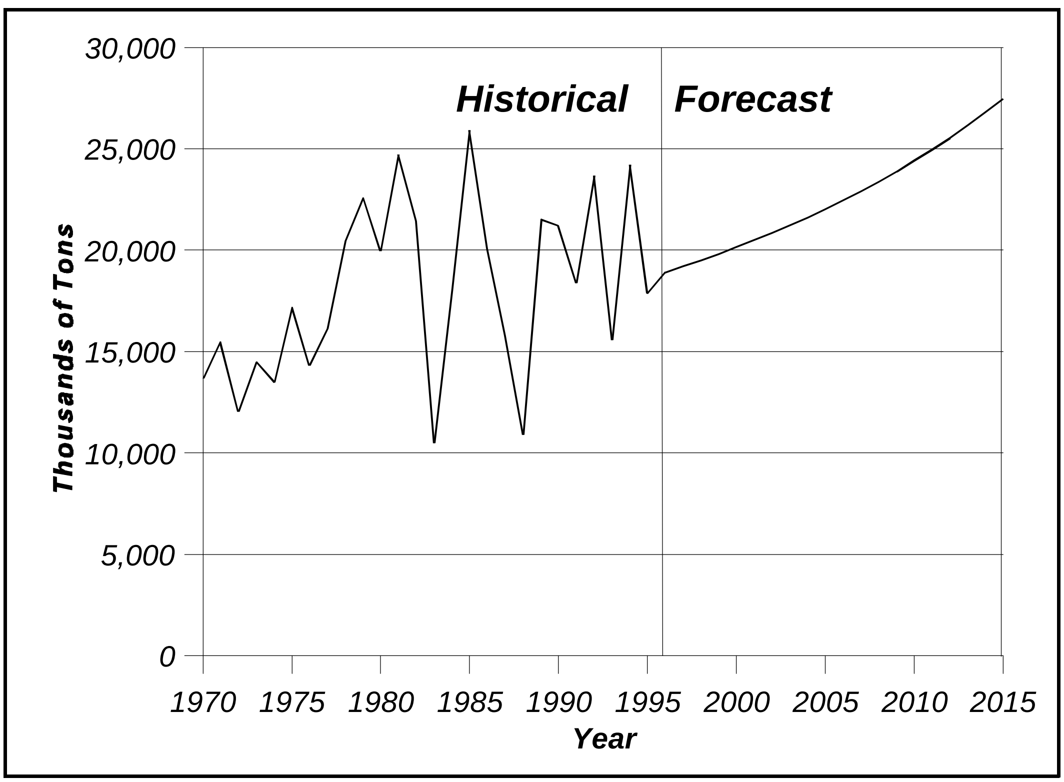

FIGURE III-1
FORECAST OF TOTAL GRAIN SHIPMENTS, CHADWICK LOCK

TABLE III-2 FORECAST GRAIN SHIPMENTS CHADWICK LOCK (1996-2015) (THOUSANDS OF TONS)					
Forecast Year	**Corn**	**Wheat**	**Soybeans**	**Other Grains**	**Total of all Grains**
1996	15,512	1,040	2,218	153	18,923
1997	15,608	1,109	2,345	151	19,213
1998	15,705	1,182	2,479	149	19,515
1999	15,802	1,260	2,621	147	19,830
2000	15,900	1,343	2,771	145	20,159
2001	15,999	1,432	2,929	143	20,503
2002	16,098	1,527	3,097	141	20,863
2003	16,198	1,628	3,274	139	21,239
2004	16,298	1,735	3,461	137	21,632
2005	16,399	1,850	3,659	136	22,044
2006	16,501	1,972	3,868	134	22,476
2007	16,603	2,103	4,090	132	22,928
2008	16,706	2,242	4,324	130	23,402
2009	16,810	2,390	4,571	128	23,899
2010	16,914	2,548	4,832	127	24,421
2011	17,019	2,716	5,109	125	24,969
2012	17,124	2,896	5,401	123	25,545
2013	17,231	3,087	5,710	122	26,150
2014	17,337	3,292	6,037	120	26,786
2015	17,445	3,509	6,382	118	27,454

SOURCES OF UNCERTAINTY

There are two direct sources of uncertainty in this forecasting method as just applied. The assumptions for the base amounts of shipments (B_i) and the growth rates (G_i) are subject to error. It is quite difficult to chose a "representative" base year from a set of observations that shows year-to-year variation. The selection of the average of shipments over the last ten periods is already an attempt to account for this variation. Secondly, assuming a growth rate as fixed over time is a dubious proposition. Economic conditions can change, drought can occur, and so can agricultural trade policies.

This approach admits a degree of ignorance concerning the factors that cause variation in commodity shipments over time. For example, good weather conditions can increase grain yields, which increase the amount of grains produced. Further, lower costs of barge shipment may increase the amount of commodities shipped over the waterway for any given level of grain production. The point is that this method uses information only on the effects of inherently complex causal relationships.[1]

ACCOUNTING FOR UNCERTAINTY

The sections below incorporate uncertainty into the forecast developed above. Because the forecasting approach of using growth rates is naive and simple, so too will be the procedures that will be used to incorporate risk and uncertainty. As a small step toward improving the information provided by this type of forecasting approach, the following sections use data on the historical variation of grain shipments in order to estimate the degree variation or uncertainty around the point forecasts shown in Table III-2.

Portraying Uncertainty in Base Shipment Levels

As mentioned before, it is difficult and inherently subjective to select the base shipment amount to which to apply the commodity-specific growth rates. If one chooses an unordinarily high amount, then forecasts of future shipments may be substantially overstated. Even worse, one may be accused of biasing the forecast by selecting a "convenient" base level of tonnage. The selection of the ten-year annual average level of shipments for the parameter B already represents an attempt to reduce these risks. However, it is further assumed here that the values of the B_i may vary in accordance with the variation in grain shipments over the 1986-1995 period.

The standard deviation (σ) is a measure of variation that can be used readily to reflect uncertainty in the assumptions for the B_i.[2] Analysis of the data reported in Table III-1 reveals that the standard deviation of annual grain shipments over the 1986-1995 period is as follows:

---

[1]However, a principal reason for selecting the "growth rate" method for forecasting is to avoid the cost of data collection and statistical estimation necessary for portraying these complex relationships.

[2]The standard deviation, σ, of a set of n observations on variable x is defined as the square root of variance, σ^2, which is defined as :

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

where μ denotes the average of the n observations on x.

Crop	Standard Deviation (1000 tons)
corn	3,722
wheat	333
soybeans	353
other	50

If the values of "true" base shipment level of each crop are normally distributed around its assumed (mean) value B with a standard deviation as defined above, then one arrives at the following well-known results:[3]

(1) Approximately 68 percent of all possible values of B will lie within the interval $B \pm \sigma$.

(2) Approximately 95 percent of all possible values of B will lie within the interval $B \pm 2\sigma$.

(3) Approximately 99 percent of all possible values of B will lie within the interval $B \pm 3\sigma$.

Further, one may define a 90 percent *confidence interval* on base shipment levels as $B \pm 1.645\sigma$. Figure III-2 illustrates an assumed distribution for base shipments of corn and shows the symmetrical nature of the normal distribution about its mean. Aside from a desired simplification, why might one assume a normal distribution for grain shipments? One very plausible reason is that weather affects grain production, and weather variables (such as rainfall and cooling degree days) are commonly assumed to be normally-distributed variables. Furthermore, if one believes that movements of grain past Chadwick Lock are reflective of steady trends in technological change and grain demand, then one might believe that variation on either side of this trend could be considered random.

Portraying Uncertainty in Future Grain Shipments by Crop

The results above may easily be extended to all forecast periods by increasing base shipments (B) <u>and</u> the standard deviation of shipments of each crop annually by the assumed long-term rates of growth. Thus, it is assumed that as volume of shipments grow, so does its variation. Under this assumption, variation in shipments grows nominally, but as a *fixed percent* of shipments. This is equivalent is to assuming a constant *coefficient of variation* in grain shipments.[4]

[3]These results reflect the standard properties of the Normal Distribution and thus of any *normally-distributed* random variable.

[4]The reader should be warned about using this method in cases of negative trend. When there is a negative trend, such as in the "other grains" category, the assumption of a constant coefficient of variation makes the forecast variance estimate decrease with time. This counter-intuitive and unlikely result suggests that one is becoming more certain about the forecast as it nears the forecast horizon. Although not done here for the sake of clarity, an option could be to fix variance at a certain *number* instead of as a percent of the mean. This would imply an increasing coefficient of variation over time just as in the positive trend case.

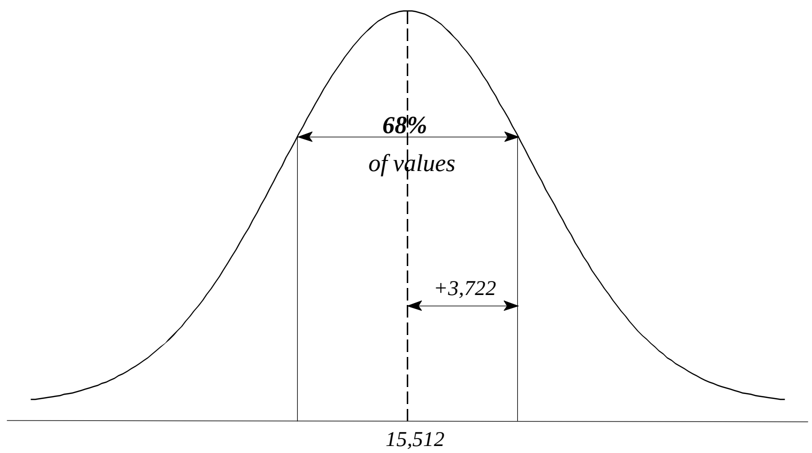

FIGURE III-2

ILLUSTRATION OF NORMAL DISTRIBUTION AROUND BASE SHIPMENTS OF CORN

Under the assumption of normality, the interval forecast of shipments for a particular grain and forecast year may then be generalized as:

$$VS_{i,f} \;=\; [\, B_i * (1 + G_i)^f \,] \;\pm\; z_c * [\, \sigma_i * (1 + G_i)^f \,] \qquad \textbf{(3.4)}$$

where z_c denotes the number of standard deviations from the mean of a normally-distributed variable for a given level of confidence, c (Note: c = 1.645 for the 90 percent confidence level). For example, the 90 percent confidence interval on forecast corn shipments in forecast period 5 (i.e., f=5 or the year 2001) is:

$$\begin{aligned} VS_{corn,\,5} &= [\, 15{,}512 * (1 + 0.0062)^5 \,] \;\pm\; z_{0.90} * [\, 3{,}722 * (1 + 0.0062)^5 \,] \\ &= 15{,}999 \;\pm\; (1.645 * 3{,}839) \\ &= 15{,}999 \;\pm\; 6{,}315 \qquad \textbf{(3.5)} \end{aligned}$$

$$or$$

$$9{,}684 \;\leq\; VS_{corn,\,5} \;\leq\; 22{,}314$$

Table III-3 presents an interval forecast of shipments by crop using this technique. The lower and upper bounds provided in the table represent the 90 percent confidence interval on the forecasts of shipments by crop.

TABLE III-3 FORECAST OF INDIVIDUAL CROP SHIPMENTS WITH UNCERTAINTY (THOUSAND OF TONS)																
	Total Quantity of Corn				**Total Quantity of Wheat**				**Total Quantity of Soybeans**				**Total Quantity of Other Grains**			
Year	**Mean**	**Std. Deviation**	**Lower Bound**	**Upper Bound**	**Mean**	**Std. Deviation**	**Lower Bound**	**Upper Bound**	**Mean**	**Std. Deviation**	**Lower Bound**	**Upper Bound**	**Mean**	**Std. Deviation**	**Lower Bound**	**Upper Bound**
1996	15,512	3,722	9,389	21,635	1,040	333	492	1,588	2,218	353	1,637	2,799	153	50	71	235
1997	15,608	3,745	9,448	21,769	1,109	355	525	1,693	2,345	373	1,731	2,959	151	49	70	232
1998	15,705	3,768	9,506	21,904	1,182	378	559	1,805	2,479	395	1,830	3,128	149	49	69	229
1999	15,802	3,792	9,565	22,040	1,260	403	596	1,924	2,621	417	1,935	3,307	147	48	68	226
2000	15,900	3,815	9,624	22,176	1,343	430	636	2,051	2,771	441	2,045	3,496	145	47	67	223
2001	15,999	3,839	9,684	22,314	1,432	459	678	2,187	2,929	466	2,162	3,696	143	47	66	220
2002	16,098	3,863	9,744	22,452	1,527	489	723	2,331	3,097	493	2,286	3,907	141	46	65	217
2003	16,198	3,887	9,804	22,591	1,628	521	770	2,485	3,274	521	2,417	4,131	139	45	64	214
2004	16,298	3,911	9,865	22,731	1,735	556	821	2,650	3,461	551	2,555	4,367	137	45	64	211
2005	16,399	3,935	9,926	22,872	1,850	592	876	2,825	3,659	582	2,701	4,617	136	44	63	208
2006	16,501	3,959	9,988	23,014	1,972	632	934	3,011	3,868	616	2,856	4,881	134	44	62	206
2007	16,603	3,984	10,050	23,157	2,103	673	995	3,210	4,090	651	3,019	5,160	132	43	61	203
2008	16,706	4,009	10,112	23,300	2,242	718	1,061	3,423	4,324	688	3,192	5,456	130	43	60	200
2009	16,810	4,033	10,175	23,445	2,390	765	1,131	3,649	4,571	727	3,374	5,768	128	42	59	197
2010	16,914	4,058	10,238	23,590	2,548	816	1,206	3,890	4,832	769	3,567	6,098	127	41	59	195
2011	17,019	4,084	10,301	23,736	2,716	870	1,286	4,147	5,109	813	3,771	6,446	125	41	58	192
2012	17,124	4,109	10,365	23,884	2,896	927	1,371	4,421	5,401	860	3,987	6,815	123	40	57	190
2013	17,231	4,134	10,430	24,032	3,087	989	1,461	4,714	5,710	909	4,215	7,205	122	40	56	187
2014	17,337	4,160	10,494	24,181	3,292	1,054	1,558	5,025	6,037	961	4,456	7,617	120	39	55	185
2015	17,445	4,186	10,559	24,331	3,509	1,124	1,661	5,357	6,382	1,016	4,711	8,053	118	39	55	182

Note: The reported means represent expected values. Upper and Lower Bounds represent 90 percent confidence intervals.

Portraying Uncertainty in Forecast of Total Grain Shipments

In order to develop a forecast of total grain shipments from the forecasts of shipments of the individual grains derived above, one must defer to some specific mathematical and statistical rules. The first of such rules is taken from Kmenta (1986):

Rule 3.1: *The expected value of a sum of random variables is equal to the sum of their expected values*:

$$E(X + Y) = E(X) + E(Y) \tag{3.6}$$

Fortunately, this rule suggests that for any particular forecast year the forecast of total grain shipments can be taken as the sum of the <u>point</u> forecasts (i.e., expected future values) of the individual grains. Given then that one can easily derive point forecasts for total grain shipments from the forecasts of individual grains, how can one derive confidence intervals around these point forecasts? Unfortunately, the answer to this question is more complicated. First, consider a very important mathematical theorem, where the *Var* denotes statistical *variance* and *Cov* denotes statistical *covariance*: [5]

Rule 3.2: *If X and Y are two random variables, then*:

$$Var(X + Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) \tag{3.7}$$

It can be shown that in the case of four random variables, such as shipments of corn (C), wheat (W), soybeans (S), and other grains (O), Equation 3.7 would expand to:

$$\begin{aligned} Var(All\ Grains) &= Var(C + W + S + O) \\ &= Var(C) + Var(W) + Var(S) + Var(O) \\ &\quad + 2\,Cov(C,W) + 2\,Cov(C,S) + 2\,Cov(C,O) \\ &\quad + 2\,Cov(W,S) + 2\,Cov(W,O) + 2\,Cov(S,O) \end{aligned} \tag{3.8}$$

Thus, if one knows the terms of Equation 3.8, then one can calculate the variance and standard deviation of the forecast values for total grain shipments. The formula for the covariance of two variables is given by:

$$Cov(X,Y) = \rho(X,Y) * \sqrt{VarX} * \sqrt{VarY} \tag{3.9}$$

---

[5] Note that variance of a variable is identical to the square of its standard deviation (c.f. footnote 2). Covariance refers to the magnitude and direction of association of two variables. For example, if large values of *X* tend to be associated with large values of *Y*, then *X* and *Y* are said to "co-vary" together and the covariance term in Equation 3.7 is non-zero and positive. As is shown below, covariance is directly related to the concept of correlation.

where ρ denotes the *coefficient of correlation*, and the latter two terms represent the standard deviations of the variables *X* and *Y*, respectively.[6] Two variables are said to be independent if their covariance is zero. It follows that independent variables are uncorrelated.

Before answering the leading question on how to derive confidence intervals on total grain shipments consider one more important statistical rule:

> Rule 3.3: *If X, Y, ..., Z are* <u>*normally*</u> *and* <u>*independently*</u> *distributed and a, b,...,c are constants, then the linear combination aX + bY + ...+ cZ is also normally distributed.* (Kmenta, 1986, <u>emphasis</u> added)

This theorem suggests that if the shipments of the individual grains are each distributed normally <u>and</u> that the shipments of individual grains are independent/uncorrelated, then one could conclude that total grain shipments is also distributed normally. Furthermore, given the mathematical rules expressed above, one could also conclude that total grain shipments would be centered around a value corresponding to the sum of the individual crops (Equation 3.6), with variance corresponding to the sum of the variance of individual grains (Equation 3.8 with covariance terms set to zero). Given that the assumption of normality has already been made for the distribution of future shipments of individual grains, a second assumption of independence would allow one to easily construct statistical confidence intervals around the point forecast of total grain shipments. Under these assumptions, the forecast interval on total shipments would be determined from:

$$VS_{total,f} = (VS_{C,f} + VS_{W,f} + VS_{S,f} + VS_{O,f}) \pm (z_c * \sqrt{Var(C)_f + Var(W)_f + Var(S)_f + Var(O)_f}) \quad \textbf{(3.10)}$$

where the term within the first set of parentheses denotes the sum of the individual commodity forecasts, the square root of the variance terms represents the standard deviation of forecasted total shipments according to application of (3.8) with independence assumption, and the z_c denotes the number of standard deviations from the mean of a normally-distributed variable for a given level of confidence, c (again, z_c = 1.645 for the 90 percent confidence level).

A Test of the Independence Assumption

Correlation analysis was undertaken using the SAS© statistical package in order to measure whether there exists any historical dependency among the levels of shipments of the individual commodities. Table III-4 presents the correlation matrix for levels of grain shipments.[7] Notice that except for the other grains group, higher growth levels of one commodity generally imply higher

---

[6] Most statistical and spreadsheet software packages provide correlation analysis routines to determine ρ. Some packages even calculate covariance as standard output.

[7] The correlations are estimated from the 26 years of annual data on the commodities reported in Table III-1.

growth rates in other commodities. Three pairs of statistically significant correlations exist: tons of corn versus tons of wheat, tons of corn versus tons of soybeans, and tons of wheat versus tons of

TABLE III-4
CORRELATION MATRIX OF ANNUAL COMMODITY SHIPMENTS*
Pearson Correlation Coefficients
(Prob> R under HO: Rho = 0 / N = 26)

Growth Rate	Corn	Wheat	Soybeans	Other Grains
Corn	1.00000	*0.57965*	*0.59257*	0.04144
	0.0	*0.0019*	*0.0014*	0.8407
Wheat		1.00000	*0.34123*	0.07830
		0.0	*0.0880*	0.7038
Soybeans			1.00000	-0.25121
			0.0	0.2158
Other Grains				1.00000
				0.0

*Correlations in italics are significant at the 10-percent level or higher.

soybeans. These findings do not support the assumption that the shipments of the individual grains are independent, which, strictly speaking, does not allow one to apply Rule 3.3 to deduce that forecast values of total shipments follow a normal distribution. Furthermore, these findings suggest that not all of the covariance terms of Equation 3.8 can be ignored.

Resolve: Assume Normality and Dependence

The findings of the correlation analysis suggest two things. First, the future shipments of the individual grains will likely be correlated. Second, because of the interdependencies, the assumption of normal distribution cannot be extended to forecast total shipments using Rule 3.3. The former finding means that variance in the forecast of total grain shipments would be better estimated if covariance is not ignored. Since covariance is calculable at low cost, it should not be ignored. The second finding does however significantly affect the ease in which one may portray uncertainty in the forecast of total grain shipments. Application of formula (3.10) is easy and consistent with how confidence intervals were placed on the forecasts of the individual grains. Keep in mind, though, that Rule 3.3 only provides a convenient means of determining whether or not a variable is normally distributed–it's a sufficient, but not a necessary, condition for assuming normality.

Aside from simplifying the analysis of uncertainty, consider why one might consider total grain shipments to be a normally distributed variable. The reasons are the same as for the individual grains. First, and foremost, grain yield is affected by weather. Weather variables (such as rainfall and cooling degree days) are commonly assumed to be normally-distributed variables. It is conceivable that the change in grain production and shipments past Chadwick Lock over time is determined by technological change and grain demand. Variation on either side of this trend could be considered random deviations. Considering the simplicity of the forecasting with growth rates methodology, these arguments are probably enough to establish the normality assumption for total grain shipments. Here, this simplification will be offset to some degree by incorporating covariance of the individual crops into formula (3.10) by using the correlation coefficients of Table III-4 in conjunction with the formula of (3.9). Thus, the formula that will be used here to establish a 90 percent confidence interval on total grain shipments for a forecast period *f* is:

$$VS_{total,f} = (VS_{C,f} + VS_{W,f} + VS_{S,f} + VS_{O,f}) \pm (1.645 * \sigma_f) \qquad \textbf{(3.11)}$$

where the first term in parentheses reflects the sum of the point forecasts of individual crop shipments (see Table III-3) and σ_f is the square root of the term calculated from application of Equation (3.8) above.

Table III-5 reports the results of the commodity forecast for total grain shipments. For comparison, the table includes forecast intervals based on dependence and independence of individual grain shipments. These results are illustrated graphically in Figure III-3. The middle forecast line connects the expected values (averages), while the top and bottom lines reflect the envelope within which 90 percent of future commodity tonnages would be expected to fall, given the assumptions that were made for the analysis. The figure shows that the forecast confidence intervals widen with the assumed steady growth in variance over time. The diagram also shows that incorporating correlation among the expected shipments of individual crops widens the confidence bands. Finally, a comparison of this figure, which incorporates and portrays an analysis of uncertainty, with Figure III-1 that did not, clearly shows the potential risk that is involved in planning based on point estimates of future barge shipments.

SUMMARY

This chapter incorporated uncertainty analysis into a forecasting methodology that is founded on applying fixed commodity-specific growth rates to base commodity shipment levels. As explained, this approach to forecasting naively ignores the factors that cause variation in commodity shipments over time. As such, the growth rate approach is simple and basic. It was shown that applying statistical confidence intervals to point forecasts developed through the use of this method was not as simple, but still basic.

The Normal Distribution and the assumption of normality were introduced as convenient and powerful tools for uncertainty analysis. The confidence intervals of this chapter were developed strictly under the assumption that future shipments of grains are distributed normally around an assumed trend that was derived from the historical 26-year rate of growth. Some established statistical rules were utilized in aggregating the forecasts of the individual commodities into a forecast of total grain shipments. These rules are instrumental to developing an understanding of mathematical expectations and variance within this and other forecasting contexts.

TABLE III-5 FORECAST OF TOTAL GRAIN SHIPMENTS WITH UNCERTAINTY							
		Dependence Among Individual Grains			Independence Among Individual Grains		
Forecast Year	**Point Prediction**	**Std. Deviation**	**Lower Bound**	**Upper Bound**	**Std. Deviation**	**Lower Bound**	**Upper Bound**
1996	18,923	4,144	12,106	25,740	3,754	12,748	25,098
1997	19,213	4,194	12,313	26,112	3,781	12,994	25,432
1998	19,515	4,246	12,531	26,499	3,808	13,251	25,779
1999	19,830	4,300	12,757	26,903	3,836	13,520	26,141
2000	20,159	4,356	12,995	27,324	3,865	13,802	26,517
2001	20,503	4,414	13,243	27,764	3,894	14,097	26,910
2002	20,863	4,474	13,502	28,223	3,925	14,407	27,319
2003	21,239	4,538	13,774	28,703	3,956	14,731	27,747
2004	21,632	4,604	14,059	29,205	3,988	15,071	28,193
2005	22,044	4,673	14,357	29,731	4,022	15,428	28,660
2006	22,476	4,745	14,670	30,281	4,057	15,803	29,149
2007	22,928	4,820	14,998	30,857	4,093	16,195	29,660
2008	23,402	4,900	15,342	31,462	4,130	16,608	30,196
2009	23,899	4,982	15,703	32,095	4,170	17,040	30,758
2010	24,421	5,069	16,082	32,760	4,211	17,495	31,348
2011	24,969	5,161	16,480	33,459	4,254	17,972	31,967
2012	25,545	5,257	16,898	34,192	4,299	18,473	32,617
2013	26,150	5,358	17,336	34,963	4,347	18,999	33,301
2014	26,786	5,464	17,798	35,773	4,398	19,551	34,020
2015	27,454	5,576	18,283	36,626	4,452	20,131	34,777

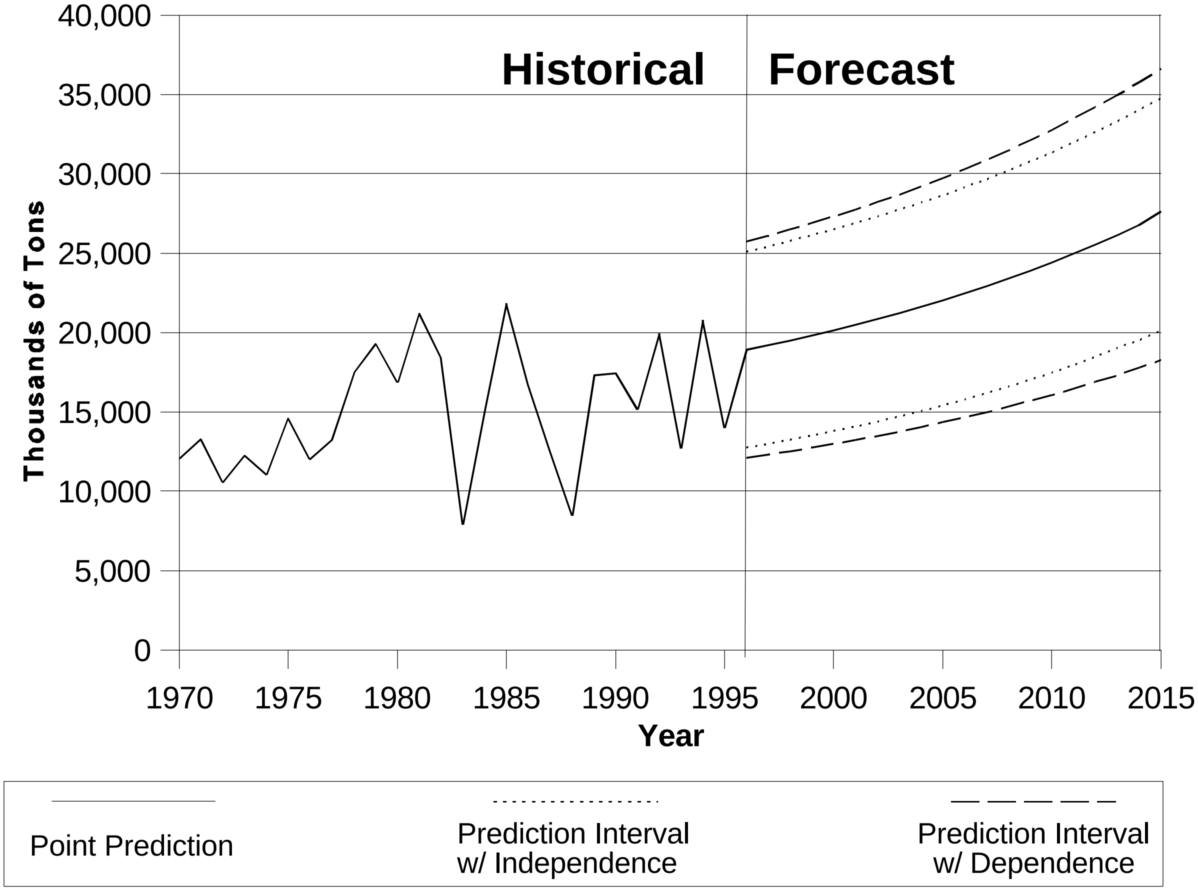

FIGURE III-3

FORECAST OF TOTAL GRAIN SHIPMENTS WITH UNCERTAINTY

It is important to keep in mind that the analysis ignored, both for simplicity and lack of data, the potential for uncertainty in the long-term rates of growth of the individual commodities[8]. Still, the effort that was undertaken to account-for and portray uncertainty should be considered a marked improvement over the customary presentation of simple point predictions.

[8]Note that in this method, a moderate swing in growth rates can lead to dramatic growth or decline in the forecasted variable.

IV. FORECASTING USING SHIPPER SURVEYS

One method that is often used to forecast the volume of commodities transported via the waterway, is simply to ask existing and potential waterway users about their plans to ship by barge. This method, also known as the "shipper survey" method, relies on formal interviews regarding current expectations of waterway use. The general procedure to conduct the "shipper survey" is rather simple:

(1) Identify all shippers in the waterway hinterland.

(2) Survey these shippers via telephone, mail, or personal interview with respect to the expected volume of a commodity to be shipped during a particular of time (say, during a specified year in the future).

EXAMPLE APPLICATION

Consider here, that twenty large shipping companies were identified within the Evanstown business economic area (BEA), who potentially could choose to transport grain to Cajun City via the Oak River. The presidents of each company were personally interviewed, with the goal of eliciting their plans to ship by barge over the next five years. Each shipper was asked to provide their best guess of how much grain, in tons, their company would expect to ship in the year 2001. Table IV-1 shows a tabulation of responses from the set of twenty shippers (S1 through S20), reported in thousands of tons, for the year 2001. The sum of their responses, 22,315,000 tons, is taken as the probable, or expected, amount of grain that would pass the Chadwick Lock in the year 2001. The idea behind this simple aggregation is that the subjective estimates of future shipments incorporate expectations of economic conditions, the relative cost to ship by barge instead of rail, and of agricultural output in the Evanstown region.

SOURCES OF UNCERTAINTY

Reliance on a shipper survey to forecast waterborne transport is visibly riddled with sources of error and uncertainty. The reliability of such a forecast is dependent on the accuracy of subjective judgements about the future. First, those who are surveyed must have a good understanding of what affects both the quantity of grains that is produced, as well as what affects their decisions to choose barge transport over rail. Uncertainty about the *ability* of those surveyed to formulate accurate expectations is a source of uncertainty that may never be surmounted. It represents risk and uncertainty at its most rudimentary level and is not quantifiable.

TABLE IV-1
EXPECTED SHIPMENTS BY SHIPPER: YEAR 2001 (THOUSANDS OF TONS)

Shippers	Expected
S1	750
S2	900
S3	900
S4	700
S5	630
S6	1,000
S7	1,000
S8	1,300
S9	1,685
S10	1,500
S11	1,950
S12	1,525
S13	1,800
S14	1,900
S15	900
S16	675
S17	875
S18	800
S19	825
S20	700
TOTAL	22,315

Second, even if one assumes that shippers are able to form accurate expectations, then the simple shipper survey identified above does not account for *how certain* the shippers are that their expectations will come true–just as in the initial application of the forecasting methods of the last chapter, one is left only with single point estimates that do not reflect any degree of uncertainty.

ACCOUNTING FOR UNCERTAINTY

The following sections describe two alternatives to the simple shipper survey that may help account for and portray uncertainty in shippers' expectations.

Shipper Survey Alternative 1

Consider that the simple survey described above was replaced with another survey that was carefully planned to assign probability to shippers' expectations. Assume that the shippers were thoroughly coached on the purpose of the survey and the need to relate their confidence in their expectations to ship by barge. Each shipper was re-interviewed and asked the following set of questions pertaining to expected barge shipment in the year 2001 by type of commodity. The purpose of the questions below was to allow each shipper to map out a cumulative probability density function around his or her expected level of shipments.

I. a. What is the amount or the volume of grain x you would expect to ship? (50th percentile)

b. What is the amount or the volume of the grain x above which you believe you will not or cannot ship? (100th percentile)

c. What is the amount or the volume of grain x below which you believe you will not ship? (0th percentile)

II. a. What is the amount or the volume of grain x above which you believe there is only a 10% chance of shipping? (90th percentile)

b. What is the amount of the volume of grain x below which you believe there is only a 10% chance of shipping? (10th percentile)

III. a. What is the amount or the volume of grain x above which you believe there is a 20% chance of shipping? (80th percentile)

b. What is the amount or the volume of grain x below which you believe there is a 20% chance of shipping? (20th percentile)

IV. a. What is the amount or the volume of grain x above which you believe there is a 30% chance of shipping? (70th percentile)

b. What is the amount or the volume of grain x below which you believe there is a 30% chance of shipping? (30th percentile)

V. a. What is the amount or the volume of grain x above which you believe there is a 40% chance of shipping? (60th percentile)

b. What is the amount or the volume of grain x below which you believe there is a 40% chance of shipping? (40th percentile)

After collecting the data from these surveys, the responses were aggregated across all shippers and all grains at the corresponding probability levels and analyzed. The results are tabulated in Table IV-2. Each row of the table represents a distribution of responses. The responses associated with the 0.5 probability level are taken to represent the means (or expected values) of these individual distributions. The standard deviation and variance reported in the last two columns of the table reflect the degree of dispersion each shipper has around his or her expected values for future shipments.

The last row of Table IV-2 reports the sums of the responses of each shipper at each probability level, as well as the corresponding variance and standard deviation of the sums. The findings suggest an expected value of shipments in the year 2001 of 22,315,000 tons, with a standard deviation of plus or minus 8,228,000 tons. This outcome requires closer inspection. By matching the lowest expectation of one shipper with the lowest expectations of other shippers, the highest expectation with other highest expectations of others, the 10th percentile with other 10th percentiles, and so on, the last row of the table in essence implies that the responses of the individual shippers are perfectly and positively correlated. Recall from the previous chapter the general formula for deriving variance of a sum of random variables X and Y:

$$Var(X + Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) \tag{4.1}$$

Perfect, positive, correlation would mean that the covariance term of this equation is at its largest, and, therefore, means that so too is the variance of the sum. Thus, the last row of the table reflects maximum variation. However, it does not seem plausible to expect that all shippers err in unison around their expected values. A more realistic scenario, for example, would be that Shipper 1 ships his expected amount (750,00 tons), while Shipper 2 ships his 70th percentile amount (1,125,000 tons), while Shipper 3 ships his 40th percentile amount (800,000 tons), and so on. In other words, the distribution of actual future shipments of the individual companies would be expected to be much less correlated. A directly opposite tact to perfect correlation would be to assume that the responses of the shippers are uncorrelated (i.e., independent), which would get rid of the covariance term of (4.1) altogether. Thus, by definition, assuming independence would comparatively reduce the anticipated amount of variation around the sum of expected shipments. Since each shipper was asked to speculate about his or her <u>own</u> plans to ship by barge, and not about plans of the group of shippers as a whole, this seems to be a fair assumption and is adopted for the present analysis.[9]

---

[9]This does, however, ignore the very real possibility that any particular shipper's expectations are based on speculation about the success of other shipping firms. For example, Shipper 1 may anticipate a chance of taking market share from Shipper 5, who may also be expecting to take market share from someone else. Under this type of scenario, the responses of the shippers would be positively, but probably not perfectly, correlated.

TABLE IV-2
DISTRIBUTION OF SHIPMENTS BY SHIPPER BY PROBABILITY LEVEL: YEAR 2001 (THOUSANDS OF TONS)

	Associated Probability												
Shipper	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**Variance**	**Std. Deviation**
S1	500	550	575	625	675	750	825	900	950	1,000	1,050	37,295	193
S2	580	600	630	700	775	900	990	1,125	1,200	1,250	1,300	75,010	274
S3	600	625	650	725	800	900	1,050	1,170	1,275	1,300	1,350	84,027	290
S4	530	540	550	600	650	700	765	825	850	900	950	23,202	152
S5	475	500	525	550	575	630	675	700	750	775	800	13,257	115
S6	625	650	725	800	850	1,000	1,125	1,260	1,350	1,450	1,475	102,636	320
S7	600	650	700	800	900	1,000	1,200	1,400	1,500	1,600	1,650	156,409	395
S8	700	750	800	950	1,050	1,300	1,450	1,800	1,900	2,000	2,050	276,409	526
S9	850	875	900	1,170	1,350	1,685	1,980	2,340	2,500	2,700	2,800	576,634	759
S10	750	825	925	1,000	1,200	1,500	1,750	2,050	2,250	2,400	2,450	429,602	655
S11	850	950	1,025	1,300	1,575	1,950	2,300	2,700	3,000	3,150	3,300	864,034	930
S12	800	825	900	1,000	1,250	1,525	1,750	2,100	2,250	2,450	2,500	444,057	666
S13	875	900	1,000	1,225	1,450	1,800	2,125	2,475	2,700	2,900	3,000	674,159	821
S14	900	950	1,000	1,275	1,500	1,900	2,200	2,600	2,850	3,050	3,150	761,011	872
S15	580	600	625	700	775	900	950	1,000	1,150	1,200	1,250	60,852	247
S16	450	465	480	540	575	675	725	800	850	900	950	33,595	183
S17	550	575	600	675	750	875	950	1,100	1,150	1,200	1,250	70,477	265
S18	475	500	550	625	700	800	925	1,050	1,175	1,225	1,250	88,011	297
S19	650	675	700	720	775	825	900	975	1,025	1,050	1,100	26,734	164
S20	600	615	625	650	680	700	775	800	850	875	900	12,205	110
Sum of Variance (S1-S20) =												4,809,620	2,193
Total	12,940	13,620	14,485	16,630	18,855	22,315	25,410	29,170	31,525	33,375	34,525	67,693,720	8,228

this assumption, the standard deviation of total expected shipments drops from 8,228,000 tons to 2,193,000 tons.

By further assuming that the distribution of forecast shipments is normal, one may readily construct a smooth and continuous distribution using the information on the mean and variance determined above.[10] An inferred forecast distribution of grain shipments in the year 2001 is represented by the Normal curve of Figure IV-1. Notice that the points of inflection pertain to ± one standard deviation from the mean. Also, from the discussion of Chapter III it is known that 90 percent of all cases would be expected to fall within ± 1.645 deviations of the mean:

$$\textit{Total Shipments} = \textit{Expected Value} \pm (z_{0.90} * \textit{Standard Deviation})$$

or

$$\textit{Total Shipments} = 22{,}315{,}000 \pm (1.645 * 2{,}193{,}000) \qquad \textbf{(4.2)}$$

or

$$18{,}708{,}000 \text{ tons} \leq \textit{Total Shipments} \leq 25{,}922{,}000 \text{ tons}$$

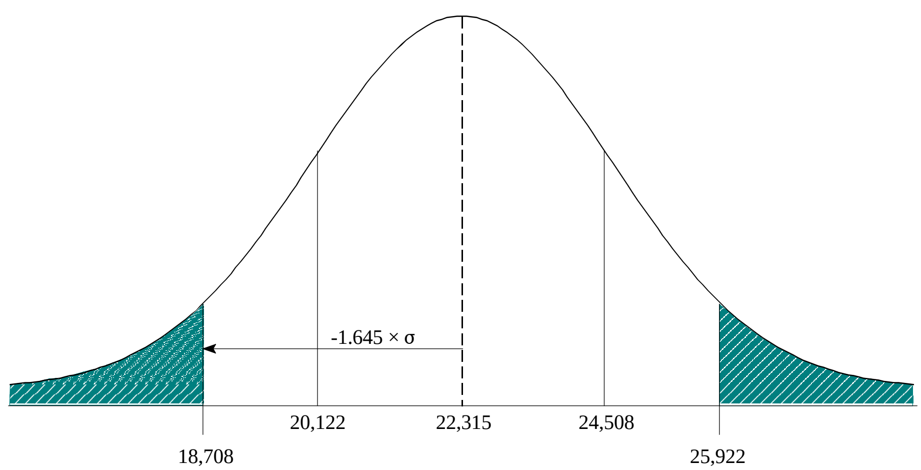

FIGURE IV-1

INFERRED FORECAST DISTRIBUTION OF GRAIN SHIPMENTS

---

[10]Recall that a normal distribution is defined by only two parameters, mean and variance. In cases like this one where the true shape of the distribution of the variable of interest is unknown, then the normality assumption is appealing and justified on the grounds of ease. If one could theorize that the responses to the shipper survey reflect 20 of very many possible responses from the same shippers (i.e., if one could suggest that the survey yielded a set of *sample averages* formed from their past experiences), then one might very loosely apply the Central Limit Theorem (CLT) to justify the assumption of normality. The CLT says that the *distribution of the sample means* is normal when the sample size is large. Here one might argue the sample size is "large" since the entire population of shippers has been sampled. As Kennedy (1992) notes, the more compelling reason to assume normality is that the normal distribution is easy to work with.

Alternative confidence intervals could be constructed by substituting different critical values for z in the above relation (e.g., 95 percent interval: $z = 1.96$; 99 percent interval: $z = 2.57$).

Shipper Survey Alternative 2

Without some expert assistance, it is likely that shippers will find it very challenging to provide the information required by alternative 1 above. It is more natural and therefore less difficult for shippers to provide most likely values together with maximum and minimum values above and below which they would not expect to ship. With this in mind, consider the following set of survey questions:

a. What is the amount or the volume of grain x you would expect to ship? (most likely)

b. What is the amount or the volume of the grain x above which you believe you will not or cannot ship? (100th percentile)

c. What is the amount or the volume of grain x below which you believe you will not ship? (0th percentile)

d. If in actuality it turns out that you ship between <*0th percentile amount*> and <*most likely amount*> of grain x, what would be your estimate of the most likely amount of grain x you would ship? (25th percentile)

e. If in actuality it turns out that you ship between <*most likely amount*> and <*100th percentile amount*> of grain x, what would be your estimate of the most likely amount of grain x you would ship? (75th percentile)

Table IV-3 reports a hypothetical set of responses for these survey questions aggregated across all grains. In this example, the answers to questions d and e above represent the midpoints (i.e., median values) between the most likely value and the minimum and maximum values, respectively. Assuming independence among the responses of the shippers, one may infer that the expected amount of total shipments in 2001 is 22,315,000 tons with a standard deviation of 2,293,000 tons. Assuming normality for total grain shipments, one may then infer the 90 percent confidence interval as: 18,543,000 tons $\leq$ total shipments $\leq$ 26,087,000 tons.[11]

Finally, one could build upon the responses to the questions above to elicit estimates of the midpoints between the 0th and the 25th percentiles (12.5th percentile) and the 75th and 100th percentiles (87.5th percentile), and so on, to map out more discrete points along the cumulative distribution. However, as with alternative 1, this likely would require facilitation from a trained interviewer.

[11] This interval is wider than the interval developed from the first survey. However, this does not mean that one should always expect higher variance from fewer questions.

TABLE IV-3
DISTRIBUTION OF SHIPMENTS BY SHIPPER BY PROBABILITY LEVEL: ALTERNATIVE 2, YEAR 2001 (THOUSANDS OF TONS)

Shipper	Associated Probability					Variance	Std. Deviation
	0	0.25	0.5	.75	1.0		
S1	500	900	750	1,260	1,050	83,570	289
S2	580	675	900	1,225	1,300	103,043	321
S3	600	825	900	1,175	1,350	87,313	295
S4	530	650	700	750	950	23,780	154
S5	475	565	630	775	800	19,093	138
S6	625	800	1,000	1,200	1,475	111,063	333
S7	600	850	1,000	1,275	1,650	163,125	404
S8	700	975	1,300	1,650	2,050	286,125	535
S9	850	1,250	1,685	2,200	2,800	592,245	770
S10	750	1,175	1,500	2,225	2,450	506,063	711
S11	850	1,400	1,950	2,600	3,300	933,250	966
S12	800	1,150	1,525	2,000	2,500	454,500	674
S13	875	1,200	1,800	2,500	3,000	779,375	883
S14	900	1,350	1,900	2,250	3,150	746,750	864
S15	580	725	900	1,050	1,250	69,505	264
S16	450	600	675	800	950	36,375	191
S17	550	990	875	1,450	1,250	120,145	347
S18	475	625	800	1,000	1,250	93,563	306
S19	650	725	825	975	1,100	33,563	183
S20	600	650	700	800	900	14,500	120
					Sum of Variance (S1-S20) =	5,256,943	2,293
Total	12,940	18,080	22,315	29,160	34,525	73,958,592.5	8,600

SUMMARY

A shippers survey elicits the expectations of those who use the waterway. The analyst must transform these expectations into forecasts of commodity flows. This chapter has shown that in order better to accommodate uncertainty analysis, a shipper survey must be designed to elicit the degree of variation (or uncertainty) that respondents might have around their expectations of the future.[12] With the aid of standard formulae, this information can be translated into uncertainty about the total amount of commodities that will move on the waterway.

Finally, it is risky to assume that shippers would be able to form accurate expectations about the distant future. As in this chapter, this technique should be adopted only to forecast commodity movements in the near-term.

[12]For more information and explanation on subjective probability exercises and expert elicitation, one may refer to the following Corps reports: *Expert Elicitation of Unsatisfactory-Performance Probabilities and Consequences for Civil Works Facilities* (Ayyub et al, 1996); *POE Lock System Risk Analysis* (Beim and Hobbs, 1994).

V. FORECASTING USING REGRESSION ANALYSIS

Regression analysis is usually used to estimate a direct and quantifiable numeric relationship between a variable of interest (the dependent variable) and a set of independent variables that are hypothesized to affect or *explain* changes in the variable of interest. The general linear regression model may be expressed as:

$$Y = \alpha + \sum_{m} \beta_m X_m + \varepsilon \tag{5.1}$$

where

Y	=	the dependent variable of interest
X	=	the mth explanatory variable
α	=	unknown model intercept term
β	=	unknown model parameters that measure the relationship between X's and Y
ε	=	stochastic disturbance (or error) term

Historic observations on Y and the vector of X's are assembled to estimate the regression equation. Linear regression selects values for α and β_m , $\hat{\alpha}$ and $\hat{\beta}_m$ that best explain changes in Y, or in statistical terms those values of α and β_m that minimize the sum of squared errors. The regression model then can be used to forecast unknown future values of Y given (or, conditioned on) future values of the explanatory variables, X_m .

In general, the *forecasted* value of Y will differ from the *true* future value of Y for any one or combination of the following reasons:

> <u>Random Error</u>: The presence of the disturbance term in equation (5.1) indicates that the estimated relationship between X_m and Y is not mathematically precise (Kexel, 1988). Forecasted values of Y implicitly assume the regression error term is zero, since its expected value is zero (i.e., $E(\varepsilon)=0$), when in fact it may differ considerably from zero due to the stochastic character of the process described (Kennedy, 1992).
>
> <u>Sampling Error</u>: Equation (5.1) is typically estimated from a sample of data, and not from data for the entire population of pairs of X_m and Y. Thus, the values of the parameters α and β_m are determined from sample data. For any given sample, the estimates of these parameters may differ from the true underlying values of the parameters (Kocik et al., 1993). In other words, the sample regression line may not exactly be the same as the population regression line (Kmenta, 1986).
>
> <u>Conditioning Error</u>: Forecasts of the dependent variable Y, are determined by, or conditioned on, presumed future values for X_m, which may be inaccurate. If assumed future

values of X_m are not realized, a discrepancy will exist between the actual value of Y and its forecasted value.

Specification Error: Errors will be introduced if the chosen regression model does not accurately represent the factors that cause changes in Y. Although theory and experience may together help one understand and build a causal model for the dependent variable, equation (5.1) may not adequately model the "real world" and the processes that generate Y (Kennedy, 1992). For example, the model may not include all relevant independent variables, the functional form of the model may be incorrect (e.g., Y may be nonlinear with respect to some or all of the independent variables), or the model parameters may change over time.

Together, these four sources of error comprise the degree of uncertainty that is inherent in forecasting using the regression approach.

The following sections apply regression techniques to model the annual number of barge tows passing the Chadwick Lock. Two examples of using regression analyses to forecast waterway traffic are discussed, one based on trend regression, and another that models the number of tows as a function of socioeconomic phenomena. The sources of uncertainty in each approach are defined, as are ways of incorporating and portraying this uncertainty in the forecast of barge traffic.

FORECASTING USING TREND REGRESSION

Consider that analysis of the historic annual number of barge tows passing the Chadwick Lock indicates a steady upward trend in barge shipments. To quantify this relationship, regression analysis is used with time as the only explanatory variable:

$$TOWS = \alpha + \beta \; Year + \varepsilon \tag{5.2}$$

where

TOWS	=	number of tows passing Chadwick Lock each year
Year	=	variable for time, measured as calendar year (1980, 1981,..., etc.)
α	=	unknown intercept term
β	=	unknown parameter that measures change in TOWS given change in Year
ε	=	stochastic disturbance term

Table V-1 lists the data that were used to estimate the parameters of the regression procedure–namely two columns of data, one column for the number of barge tows and one column

TABLE V-1 HISTORICAL OBSERVATIONS FOR THE NUMBER OF TOWS PASSING THE CHADWICK LOCK	
Year	**Tows**
1970	802
1971	887
1972	704
1973	817
1974	736
1975	975
1976	800
1977	882
1978	1,162
1979	1,286
1980	1,121
1981	1,415
1982	1,227
1983	526
1984	1,003
1985	1,457
1986	1,115
1987	839
1988	564
1989	1,156
1990	1,163
1991	1,010
1992	1,330
1993	847
1994	1,385
1995	932

that designates the corresponding year. As shown, the variable *Year* takes on values from 1970 to 1995. Table V-2 describes the estimated parameters of the regression equation (i.e., $\hat{\alpha}$ and $\hat{\beta}_{Year}$).[12] The coefficient for the time variable is indeed positive and statistically significant at the 0.10 level.[13] The R-squared value indicates that the regression relationship explains only 7 percent of the variation in the annual number of barge tows. Ignoring the error term of (5.2) for the time being, the following equation can now be used to forecast future barge traffic passing the Chadwick Lock:

$$TOWS = -20{,}481 + (10.837949 * Year) \qquad (5.3)$$

Figure V-1 illustrates the outcome of using this equation to predict barge tows over a 20-year forecast horizon by substituting particular years for the variable *Year*. As in the initial growth rate example of Chapter III, the forecast line in smooth and not indicative of the variation that occurred during the historical period.

TABLE V-2
TREND REGRESSION ANALYSIS

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Prob > T
Intercept ($\hat{\alpha}$)	-20,481	12,830.4798	-1.596	0.0135
Year ($\hat{\beta}_{\text{Year}}$)	10.837949	6.4718	1.675	0.1070

Dependent Variable: Tows
N = 26
Adj. R^2 = 0.0673
F-Value = 2.804
Prob>F= 0.1070
Root MSE = 91.7005

[12] The regression statistics that are shown represent just some of the standard output of the regression procedure.

[13] Without going into great detail, statistical significance of parameters in the regression model is determined by comparing the magnitude of the coefficient estimate with its standard error. The *t*-value corresponds to the ratio of the parameter estimate to its standard error. Thus, the higher the *t*-value, the higher the confidence that one may place in the reliability of the parameter estimate. The reader is referred to the following texts for a more comprehensive treatment of the meaning, interpretation, and caveats in the use of *t*-values for statistical inference, as well as for other special topics on interpreting regression results: Kennedy (1992), Kmenta (1986), Judge et. al. (1988).

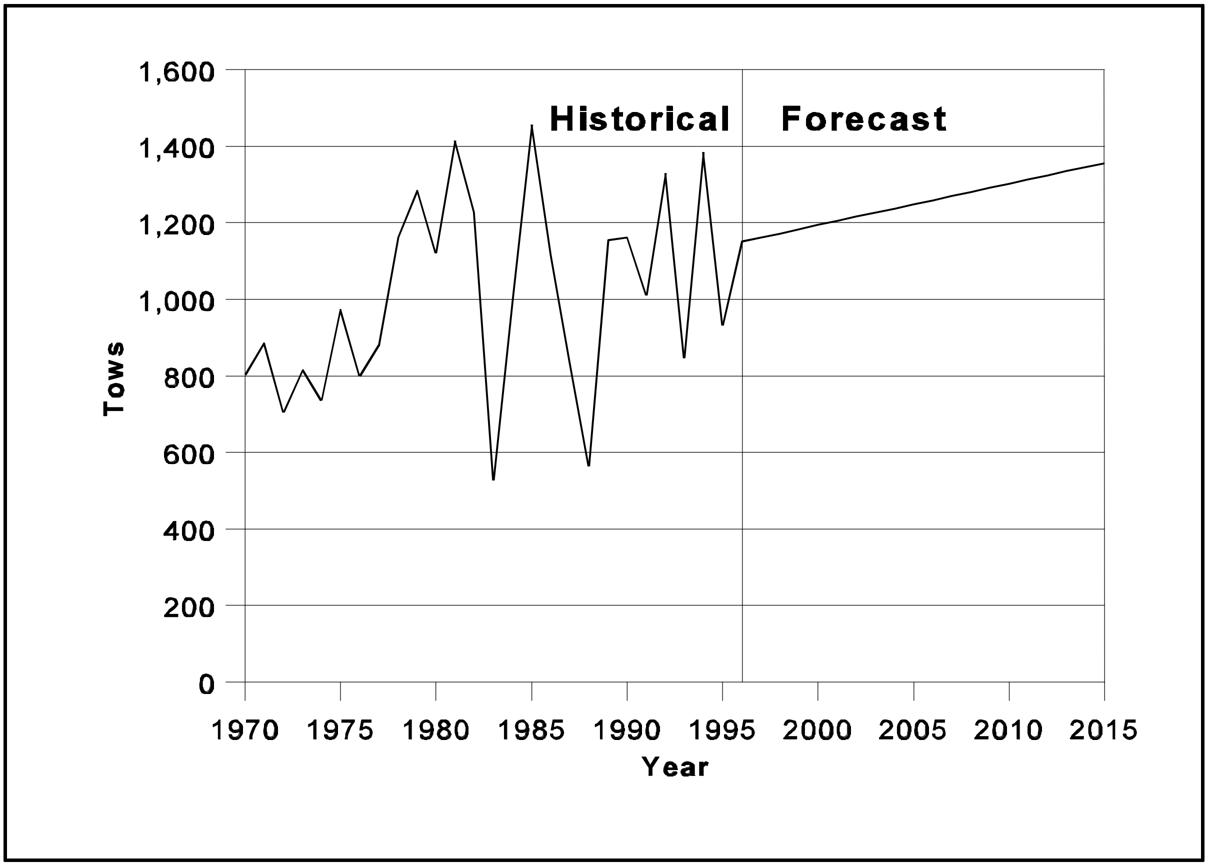

FIGURE V-1
HISTORICAL AND FORECASTED VALUES OF TOWS

Sources of Uncertainty

Within the context of using regression analysis to forecast tows using time as a single regressor, there are only three of the four sources of uncertainty (or error) as described above. There is no conditioning error on the future values of the time variable, since future values of the time variable are certain. There is, though, sampling error, since within the historic data, the time variable represents only a sample (1970, 1971,...,1995) of possible values.[14] Also, the specification of time as the only independent variable in the model blatantly ignores the factors that cause changes in tows

[14] It is also interesting to note that there is no *measurement error* on the time variable. For example, the year 1980 is plainly 1980, and not 1980.34. For other, less discrete, independent and dependent variables, measurement errors may lead to biased regression parameters, particularly if the errors in measurement are not random (i.e., if the errors are systematic). More sophisticated statistical techniques can be used to cope with errors in measurement, but are beyond the scope of this manual. The analyses in this and all other chapters presume that variables are measured without error.

and represents a strong possibility for specification error.[15] Furthermore, with time as the only independent variable, the model assumes implicitly that the same trend in barge traffic will occur until the end of time.

Accounting for Uncertainty

The most immediate way to reduce specification error would be to re-estimate the regression model with causal factors defined on the right-hand side. This is done later in this chapter. This section will focus on accounting for and portraying the random and sampling error that is inherent in equation (5.2), using standard procedures.

Regression analysis uses standard formulae to construct prediction intervals around point forecasts that account for random and sampling error. A prediction interval for the future number of tows, $TOWS_f$, is constructed as:

$$TOWS_f = TOWS_p \pm (s_f * t_c) \qquad (5.4)$$

where $TOWS_p$ represents the point prediction for future tows estimated from the regression relationship (5.3), s_f denotes the standard forecast error, and t_c is the value of the standard t-statistic for a given level of confidence, c.

The standard forecast error is calculated as the square-root of forecast error variance, which in the case of a single explanatory variable (here, time measured by *Year*), may be expressed as:

$$s_f^2 = s_m^2 \left[1 + \frac{1}{N} + \frac{(Year_f - \overline{Year})^2}{\sum (Year_A - \overline{Year})^2} \right] \qquad (5.5)$$

where $\overline{Year}$ represents the sample mean of the time variable (here 1982.5), $Year_A$ represents a specific year of the historical data (e.g., 1990), $Year_f$ represents the numeric value of the forecast year (e.g., 2005), and

$$s_m^2 = \frac{1}{N-2} \Sigma \, (TOWS_A - TOWS_p)^2 \qquad (5.6)$$

[15]Time is the general context within which particular factors operate to bring about movements in tows through the Chadwick Lock. Thus, to a large degree, this specification error would be expected to bring about the so-called random error of the regression equation.

denotes the mean variance (or squared error) of the discrepancies between the actual historical number of tows and the number of tows predicted from the regression model.

The formulae above provide an understanding of the components of forecast error within the context of regression. Equation (5.6) illustrates that the smaller the difference between actual historical observations and those predicted by the regression relationship (i.e., the smaller the random error), the smaller the forecast error. Equation (5.5) implies that the larger the sample range and size upon which the regression equation is based, the smaller the sampling error. Also, it indicates that forecast error increases as the value of the explanatory variable (time) departs from the mean of the explanatory (time) variable in the data set used to construct the regression relationship. In other words, one is better able to forecast within the "range of experience" of the regression equation than outside of it. The "range of experience" is represented by the sample values of the explanatory (time) variable and the sample mean, which are in the denominator of (5.5) (Kmenta, 1986). Figure V-2 illustrates this important point with the characteristically widening curves of forecast confidence as the values of the independent variable (time) move away from the mean of the data.

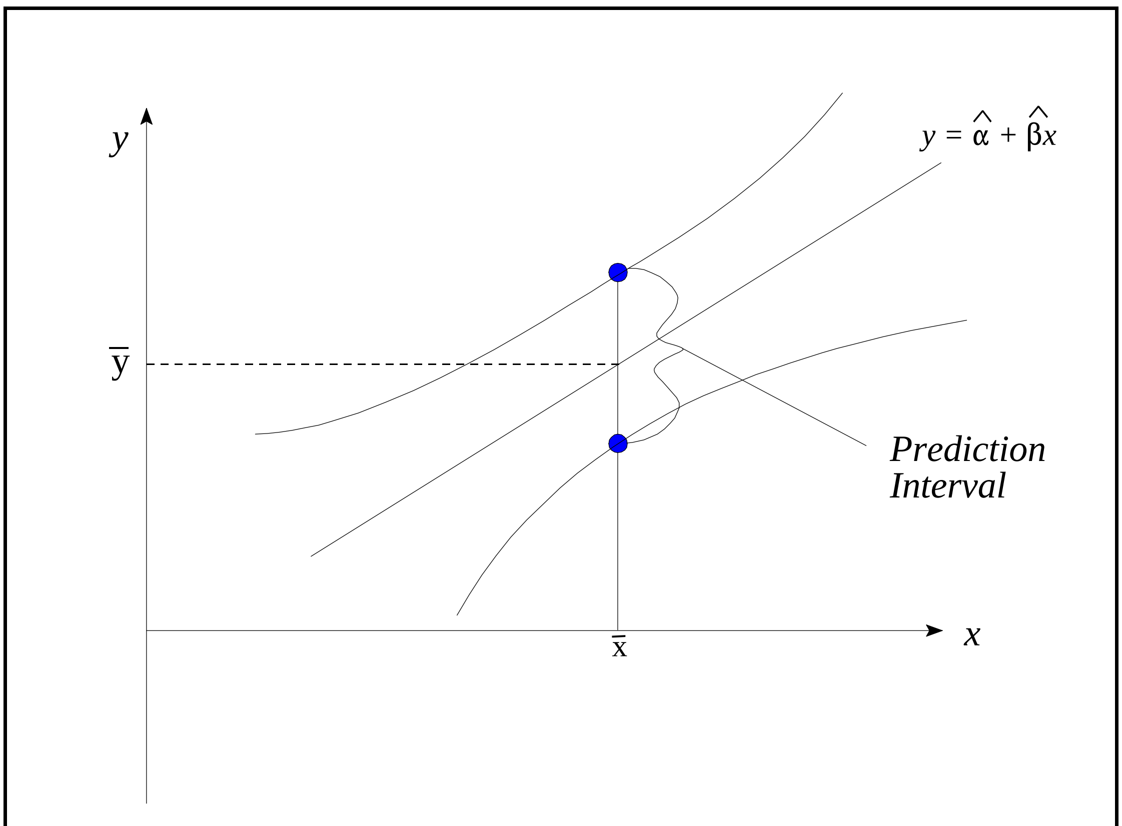

FIGURE V-2
ILLUSTRATION OF STATISTICAL CONFIDENCE INTERVALS

The Forecast with Uncertainty

Table V-3 shows the outcome of applying the confidence interval formulae shown above to the point estimates of future tows. The first column of the table shows the values of the time variable that were substituted into the estimated regression equation. The second column presents the resultant point forecast of future barge tows. The last two columns present the lower and upper 90 percent confidence bounds for the forecast, which were derived from substitution of calculated standard forecast errors and a *t*-value of 1.714 into equation (5.3) for each forecast year.[16]

TABLE V-3
90 PERCENT CONFIDENCE INTERVALS OF FUTURE TOWS

Forecast Year	Point Prediction of Tows	Lower Bound	Upper Bound
1996	1,152	695	1,608
1997	1,162	702	1,623
1998	1,173	709	1,638
1999	1,184	715	1,653
2000	1,195	722	1,668
2001	1,206	728	1,683
2002	1,217	734	1,699
2003	1,227	740	1,715
2004	1,238	745	1,731
2005	1,249	751	1,747
2006	1,260	756	1,764
2007	1,271	761	1,780
2008	1,282	766	1,797
2009	1,292	771	1,814
2010	1,303	775	1,831
2011	1,314	779	1,849
2012	1,325	784	1,866
2013	1,336	788	1,884
2014	1,347	792	1,902
2015	1,357	796	1,919

[16]The value of the t-statistic depends on degrees of freedom and level of confidence desired. The value of 1.714 was derived from a statistical table for a sample size of 26 (n) with 24 [n-(k+1)] degrees of freedom (where k denotes the number of explanatory variables in the model, <u>excluding</u> the intercept term) and a 90 percent confidence level.

Figure V-3 presents the results graphically. The confidence bands widen slightly as time approaches the end of the forecast period in accordance with equation (5.5). The confidence bands are quite wide, which is indicative of the low explanatory power of the model and the likelihood that important variables have been omitted. Thus, specification error is almost certainly present in this simple model.

FORECASTING USING MULTIPLE REGRESSION

Recognizing some of the conceptual flaws of using time as a single independent variable, suppose that interviews with local shippers and agricultural economists pinpointed two primary determinants of barge traffic on the Oak River: the cost of shipping by barge relative to the cost of shipping by rail, and total grain production in the region. Historical data for these variables for the 1970-1995 period were obtained from various sources and matched with historical annual tow data. This data set is reported in Table V-4. Table V-5 shows the results of estimating a linear regression equation from these data. The estimated parameters for total production and relative price are statistically significant, and their size and respective signs align with prior expectations. The annual number of tows increase with higher total grain production and decrease as the cost of shipping by barge rises relative to the cost of shipping by train. The regression model explains about 87 percent of historical variation in the annual number of tows.

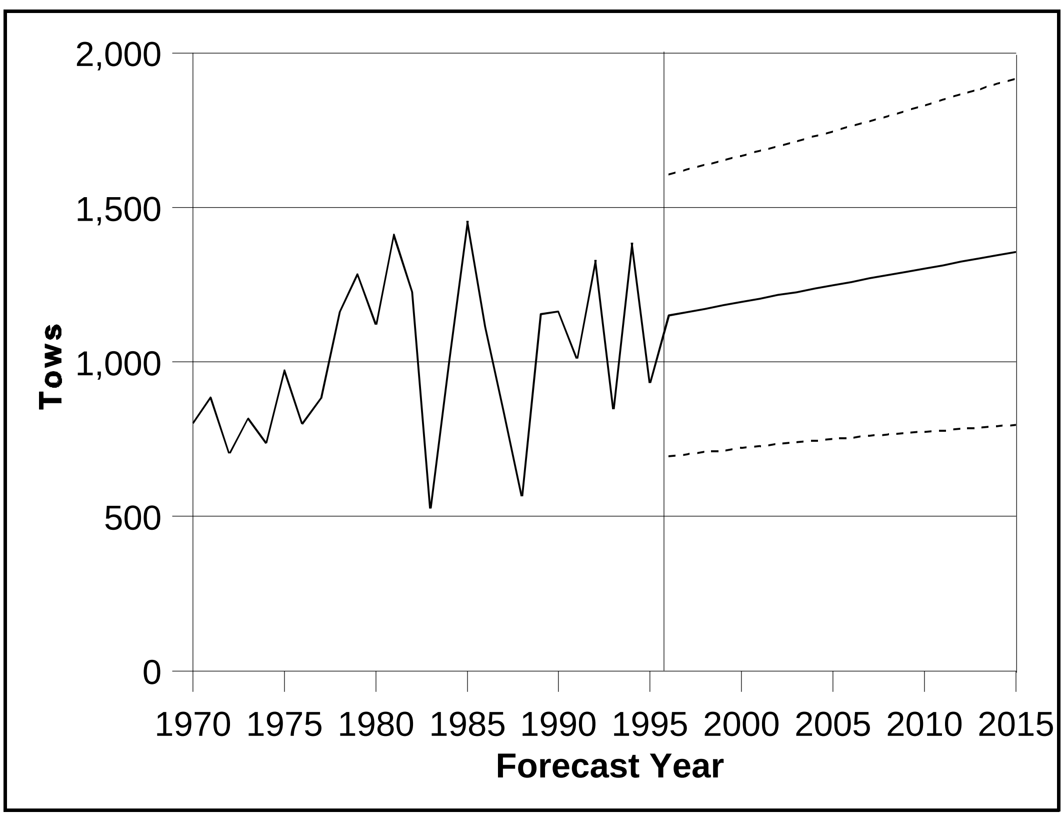

FIGURE V-3
90 PERCENT CONFIDENCE INTERVALS OF FUTURE TOWS

TABLE V-4 HISTORICAL OBSERVATIONS OF TOWS, RELATIVE PRICE, AND TOTAL PRODUCTION			
Year	**Relative Price**	**Total Production***	**Tows**
1970	0.760	37,479	802
1971	0.819	38,263	887
1972	0.838	38,269	704
1973	0.895	39,730	817
1974	0.769	31,763	736
1975	0.650	40,115	975
1976	0.820	39,020	800
1977	0.802	43,304	882
1978	0.697	47,001	1,162
1979	0.559	51,384	1,286
1980	0.751	45,231	1,121
1981	0.851	54,617	1,415
1982	0.703	53,353	1,227
1983	0.901	28,789	526
1984	0.611	47,575	1,003
1985	0.865	55,424	1,457
1986	0.896	51,653	1,115
1987	0.805	45,416	839
1988	0.802	29,111	564
1989	0.964	49,617	1,156
1990	1.079	51,654	1,163
1991	0.952	46,694	1,010
1992	0.997	58,845	1,330
1993	0.962	36,717	847
1994	0.970	63,069	1,385
1995	0.950	47,431	932

*Total production is measured in 1,000 tons.

TABLE V-5
MULTIPLE REGRESSION SPECIFICATION FOR ANNUAL TOWS
REGRESSION RESULTS

Variable	Parameter Estimates	Standard Error	T for HO: Parameter=0	Prob > T
Intercept	-36.415736	142.58999602	-0.255	0.8007
Total Production	0.027645	0.00211034	13.100	0.0001
Relative Price	-244.572843	148.66764127	-1.645	0.1135

Covariance of Estimates

	Intercept	Tot. Prod.	Rel. Price
Intercept	20331.906964	-0.152022466	-15789.30358
Tot. Prod.	-0.152022466	4.4535293E-6	-0.058373287
Rel. Price	-15789.30358	-0.058373287	22102.067562

Dependent Variable: Tows
N = 26
Adj. R^2 = 0.8720
F-Value = 86.130
Prob>F= 0.0001
Root MSE = 91.7005

The parameters of Table V-5 form the following equation used to forecast future tows moving past the Chadwick Lock:

$$TOWS = -36.4157 - (244.5728 * Relative\ Price) + (0.02765 * Total\ Production) \quad \textbf{(5.7)}$$

A forecast of tows requires substitution of data on future values of relative price and total grain production into Equation (5.7). Future values for relative price and total production were obtained from an agricultural macroeconomic forecasting model maintained at the University of Evanstown. Table V-6 reports a series of point forecasts of future tows for the years 2005, 2010, and 2015, together with the assumed future values of relative price and total production. Since the future values of relative price and total production are themselves point forecasts, they do not account for uncertainties in their predicted values and conditioning error is ignored.

Sources of Uncertainty

The point forecasts reported in Table V-6 incorporate all four sources of error defined at the beginning of this chapter. There is sampling error on relative price and total production, since the historical sampling of values for these variables reflect only one of many possible such samplings.

TABLE V-6
EXPECTED, FUTURE VALUES OF RELATIVE PRICE, TOTAL PRODUCTION, AND TOWS

Year	Expected Relative Price	Expected Total Production of Grains	Expected Tows
2005	0.90	65,000	1,540
2010	0.85	70,000	1,691
2015	0.80	75,000	1,841

Thus, the intercept term and the coefficients of relative price and total production (i.e., $\hat{\alpha}$ and the $\hat{\beta}_m$) reflect only sample *estimates* of the true values of α and β's. The model does not explain all of the historical variation in tows due to random error. Furthermore, the forecast of tows are conditioned on assumed future values for relative price and total production. Unlike the previous example that treated future values of time as certain, there is conditioning error stemming from inaccuracies in the forecasted values of relative price and total production. Finally, although the model follows a theoretical formulation of what causes changes in the number of tows passing Chadwick Lock, it may still be misspecified if other relevant explanatory variables are omitted, or if the relationship among the variables has non-linear characteristics, or if the relationship between the left- and right-hand sides of the equation has changed over time.

Accounting for Uncertainty

The sections below explore the effects of sampling, random, and conditioning error on the forecasts of tows. After presenting an updated forecast of tows that reflect these uncertainties, the chapter concludes with some remarks on how to deal with the subject of specification error.

Accounting for Random and Sampling Error

Forecast intervals for a multiple variable regression equation are constructed in the same way as in the single variable case presented above, in which equation (5.4) is used. However, the formulation of the standard forecast error requires a different formula. In the case of two independent variables, the standard forecast error, s_f, is taken as the square root of forecast error variance s_f^2, which may be calculated from the following formula:

$$s_f^2 = s_m^2 + \frac{s_m^2}{n} + (X_1 - \overline{X_1})^2 \, s_{\hat{\beta}_1}^2 + (X_2 - \overline{X_2})^2 \, s_{\hat{\beta}_2}^2 + 2\,(X_1 - \overline{X_1})\,(X_2 - \overline{X_2})\, Cov(\hat{\beta}_1, \hat{\beta}_2) \qquad (5.8)$$

where

s^2_m	=	the estimated variance of the model error term (see Equation 5.6)
X_i	=	value of the *i*th independent variable for any particular forecast period
$\overline{X}_i$	=	the mean value of the *i*th independent variable upon which the regression model was based
n	=	size of sample over which regression model was estimated
$s^2_{\hat{\beta}_i}$	=	variance of the estimated parameter ($\hat{\beta}$) of the *i*th independent variable
Cov ($\hat{\beta}_1$, $\hat{\beta}_2$)	=	covariance between estimated parameters of the *i*th and *j*th independent variables (I j)

Most statistical software packages provide s_m^2 with standard regression output. The variances of the estimated parameters represent the square of the standard errors of the coefficient estimates, and are also reported in the diagonal of the variance-covariance matrix. The variance-covariance matrix is produced as output by most statistical software packages. This matrix is shown at the bottom of Table V-5. The off-diagonal elements of this matrix represent the covariance terms.[17,18] Notice that the variance and covariance terms related to the model intercept term are not used in equation (5.8).

Although it contains more terms, the formula above operates exactly as does Equation (5.5), except that it is designed to calculate the effects of random and sampling error for a regression equation containing more than one independent variable. The first term, s_m^2, measures the effects of random error. The smaller the difference between actual historical observations and those predicted by the regression relationship (i.e., the smaller the random error), the smaller the forecast error. The

---

[17] Although tedious, the formula of equation (5.8) is easily expandable to cases of more than two independent variables. The addition of one variable brings about an addition of one more squared deviation term and a *few* more covariance terms. Equation (5.8) may be more formally expressed as:

$$s_f^2 = s_m^2 + \frac{s_m^2}{n} + \sum_k (X_k - \overline{X_k})^2 \, s^2_{\hat{\beta}_k} + 2 \sum_{j<k} (X_j - \overline{X_j})\,(X_k - \overline{X_k})\; Cov\,(\hat{\beta}_j, \hat{\beta}_k)$$

where subscripts *k* and *j* denote the *k*th and *j*th explanatory variables of the model (excluding the intercept term).

[18] The reader may be familiar with the matrix equation for forecast error variance:

$$s_f^2 = s_m^2 + X_f \, s_m^2 \, (X^T X)^{-1} \, X_f^T$$

where s^2_m denotes the variance of the model error term (see Equation 5.6), X_f is a row matrix of the forecasted values of the independent variables for any particular forecast year, and *X* is a matrix of the values of the independent variables for the sample on which the regression model was based. The superscript *T* in Equation denotes matrix transposition, while the superscript of *-1* denotes matrix inversion. The term $s^2_m\,(X^TX)^{-1}$ is the variance-covariance matrix.

second term indicates that the larger the sample size upon which the regression equation is based, the smaller the sampling error. Exactly as in Equation (5.5), Equation (5.8) incorporates the fact that forecast error increases as the values of the explanatory variables depart from their respective means as measured from the data used to construct the regression relationship. Finally, the formula shows the intuitive result that the smaller the variance and covariance of the estimates of the regression parameters, the smaller the variance of forecast error.

Standard forecast errors (s_f's)were derived using Equation (5.8) through the substitution of the expected future values of relative price and total production into the equation. Next, using Equation (5.4), 90-percent confidence intervals were constructed around the point forecasts of future tows that were reported in Table V-6. Table V-7 presents the resultant interval forecast for tows. Note that the width of the intervals in these tables take into account random and sampling error only and <u>do not</u> account for the effects of conditioning error on the forecast of tows.

TABLE V-7
INTERVAL FORECASTS FOR TOWS: RANDOM AND SAMPLING ERROR ONLY

Year	Lower Bound	Expected Value	Upper Bound
2005	1,365	1,540	1,716
2010	1,507	1,691	1,874
2015	1,647	1,841	2,036

Accounting for Conditioning Error

Recall from the discussion of the point forecast, that the forecast values of relative price and total grain production were derived from external sources that did not provide any indication of the level of uncertainty associated with their forecasted values. If ranges of future values of relative price or total production had been provided, it would have been possible to substitute them into the regression model to produce alternative forecasts for tows.[19] In the absence of such data, a three step *Monte Carlo* methodology derived from Kexel (1988) is used to assess the conditioning error. The three steps involve:

(1) Selection of theoretical probability distributions for the explanatory variables (in this case the variables are relative price of barge shipment and total grain production).

(2) Analysis of correlation between explanatory variables (in this case correlation between relative price of barge shipments and total grain production).

[19] This technique is often referred to as sensitivity or scenario analysis.

(3) Monte Carlo simulation of future values of the dependent variable (i.e., number of tows), based on simulated values of the explanatory variables, where simulated values of the explanatory variables are selected from the probability distributions of step 1 and incorporate any significant correlations found in step 2.[20]

Step 1

Step 1 involved an analysis of historical data on relative price and total production, in order possibly to infer appropriate probability distributions for the explanatory variables. BestFit© probability distribution-fitting software was used to analyze the historical data. The two panels of Figure V-4 show the frequency histograms of the historical data on relative price and total production. The smooth curves in the diagram refer to the probability distributions that BestFit© selected to “best” represent the historical data. As shown, historical observations on both relative price and grain production are characterized best by the Weibull probability distribution, although BestFit© ranked the normal distribution a very close second. Figure V-4 illustrates the bell-shaped nature of the fitted Weibull distribution, and points out that the probability distributions selected by the program fit fairly well.

Despite the fit to the historical data, the selection of appropriate probability distributions for *future* values of relative price and production required further judgement to incorporate the expectations of the future noted in Table V-6. The existence of trend in the independent variables suggests that it is possible that the shape and scale of the probability distributions will change with time. Unlike the Normal distribution, which can easily accommodate such assumptions, the shape and scale of the Weibull distribution is defined by the mean and standard deviation in a very complex way. The relationship between the defining parameters of the Weibull distribution (α and β) and the defining parameters of the normal distribution (μ and σ) involves complex nonalgebraic functions. Therefore, with practical considerations in mind, future values of relative price and total production were assigned to follow Normal Distributions.

As noted, the specification of a normal distribution requires two parameters, namely a mean ( ) and a standard deviation (σ). The forecasted values for relative price and total production reported in Table V-6 were taken to represent the means of their respective normal distributions. Assumptions regarding standard deviations required an analysis of historical variation in relative price and total production. The coefficient of variation (CV) was estimated for each variable, and is simply defined as the standard deviation of a variable divided by its mean value. The historical coefficients of variation were found to be 0.1507 and 0.1963 for relative price and total production, respectively. The formula for coefficient of variation was then rearranged to derive assumptions for the standard deviations of the theoretical normal distributions:[21]

---

[20] The Monte Carlo method is a technique that is used to select randomly from a given distribution that characterizes the underlying data. More details on this technique are provided in the next chapter.

[21] This is another example of applying a constant coefficient of variation to estimate future variance. Remember that applying this technique to a variable with a downward trend will result in shrinking estimates of absolute variance over time. An option for avoiding this effect is mentioned in footnote 4.

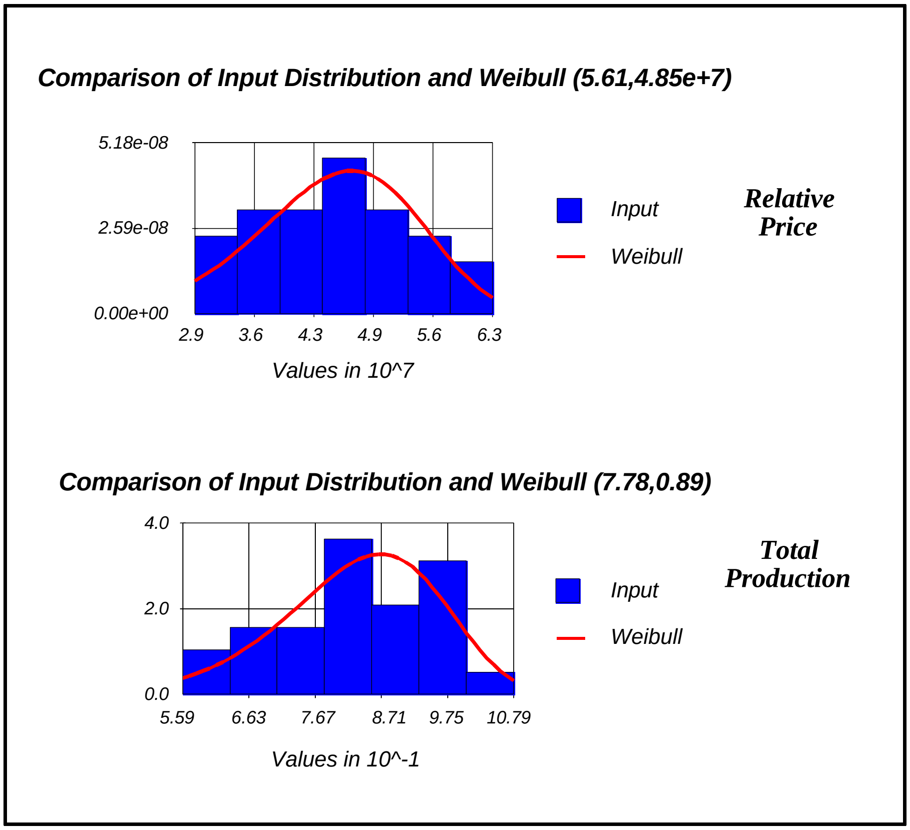

FIGURE V-4

COMPARISON OF HISTORICAL FREQUENCY DISTRIBUTIONS AND THEORETICAL PROBABILITY DISTRIBUTIONS

$$standard\ deviation_{price} = Forecasted\ Relative\ Price * CV_{price}$$

$$standard\ deviation_{tot.\,prod.} = Forecasted\ Total\ Production * CV_{tot.\,prod.}$$

Together, these assumptions defined unique theoretical normal distributions for relative price and total production for each forecast year (2005, 2010, 2015). The assigned normal distributions for the exercise are presented in Table V-8.

Step 2

For step 2, the historical data for relative price and total grain production were examined for correlation. The correlation analysis did not indicate statistically significant dependencies between the two variables. (Note, that if significant correlations had existed, they would have been incorporated into the simulations of Steps 3 and 4 using a menu driven function of the simulation software. This is demonstrated in Chapter VI.)

Step 3

The Monte Carlo simulation routine of the @Risk© software package was used to simulate a forecast of tows for each forecast year, based on iterative and random selections of values of relative price and total production from the respective normal distributions defined in Table V-8 and iterative substitution of these values into the regression model of Equation (5.7). A set of 1,500 pairs of values of relative price and total production and subsequent predictions of tows were generated by the simulation procedure for the forecast year 2005. A range of 1,600 and 1,400 values were generated for the forecast years 2010 and 2015, respectively.[22]

TABLE V-8
ASSIGNED NORMAL DISTRIBUTIONS FOR FUTURE VALUES OF RELATIVE PRICE AND TOTAL PRODUCTION

Year	Variable	Expected Value ( )	Standard Deviation (σ)
2005	Relative Price Total Production	0.90 65,000	0.13563 12,759.5
2010	Relative Price Total Production	0.85 70,000	0.128095 13,741.0
2015	Relative Price Total Production	0.80 75,000	0.12056 14,722.5

The 5th and 95th percentiles of the range of predicted tows for each forecast year were taken to construct upper and lower 90 percent confidence intervals based on conditioning error only—that is, based solely on the simulated variation in the explanatory variables. This interval forecast of tows is presented in Table V-9. Remember, however, that the intervals correspond to conditioning error only and <u>do not</u> account for the uncertainty associated with random and sampling error.

[22]The number of samplings for each forecast year was not predetermined. Rather, the simulation model for the number of tows *converged* after these number of iterations. As discussed in the next chapter, convergence occurs when a distribution becomes "stable," after which the statistics describing the distribution do not change significantly with additional iterations. Note that a simulation can converge without necessarily sampling very low probability events. In cases where such low probability events have large consequences, the simulation results should be reviewed to verify that the event(s) occurred in the simulation.

In order to incorporate random, sampling, and conditioning errors simultaneously, one must rely on Equations (5.4) and (5.8), which define the forecast interval and forecast error variance, respectively. Equation (5.8) suggests that for each pair of future values of relative price and total

TABLE V-9
INTERVAL FORECASTS FOR TOWS: CONDITIONING ERROR ONLY

Year	Lower Bound	Expected Value	Upper Bound
2005	943	1,540	2,118
2010	1,077	1,692	2,320
2015	1,179	1,842	2,504

production there exists a forecast error variance (s^2_f), and therefore, a resultant standard forecast error (s_f). Meanwhile, Equation (5.4) implies that each resultant standard forecast error must be multiplied by a value from the *t*-distribution to obtain a prediction error that may be added to the prediction of tows.

Equation (5.8) was built into the Monte Carlo simulation for each forecast year, so that simulated values of relative price and total production would generate corresponding values of s_f. Simultaneously, and using the @Risk© simulation tool, each of these values of s_f was multiplied by a corresponding value of *t*, which was randomly selected from a Student- *t* distribution defined by 23 degrees of freedom.[23] This operation resulted in a distribution of simulated values of the quantity (tows+t*s_f) for each forecast year. The 5th and 95th percentiles of the distribution of (tows+t*s_f) were then selected as the lower and upper bounds of the 90 percent confidence interval on the forecast of future tows—an interval which accounts for sampling, random, <u>and</u> conditioning error. Figure V-5 illustrates this process of accounting for these sources of error.

Table V-10 presents the forecast of future tows with uncertainty. For convenience, the table reports the 90 percent forecast interval assuming random and sampling error only, conditioning error only, and all three sources of error together. As one might expect, the intervals corresponding to all three sources of error envelop the intervals that do not consider all three sources.[24]

[23]The Student-*t* distribution is defined by only one parameter, namely, degrees of freedom. Recall that within the regression context degrees of freedom are calculated from the formula [n-(k+1)], where *n* denotes the number of observations in the data used in the regression analysis and *k* denotes the number of explanatory variables in the regression model, excluding the intercept. The Student-*t* distribution is similar to the standardized Normal Distribution, in that it is bell-shaped and symmetric around a mean of 0.

[24] Theoretically, one would expect the expected values reported in Table V-10 to match exactly. However, this is contingent on the simulated distributions matching the theoretical distributions exactly. As suggested by the results, convergence of the simulation model will likely be achieved before this result.

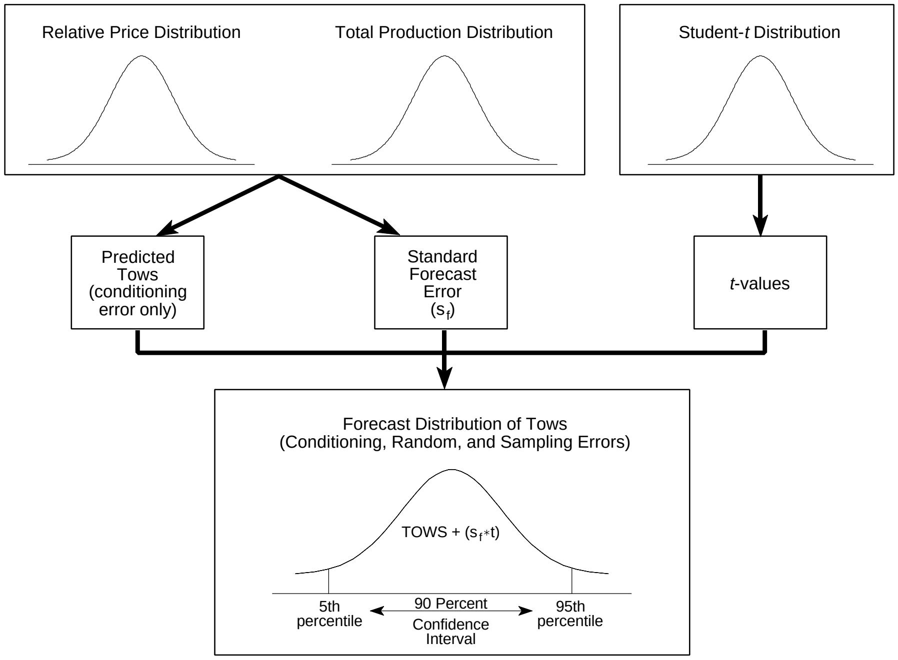

FIGURE V-5

APPLICATION OF MONTE CARLO SIMULATION WITHIN MULTIPLE REGRESSION FRAMEWORK

TABLE V-10
INTERVAL FORECAST WITH RANDOM AND SAMPLING AND CONDITIONING ERRORS COMBINED

Year	Scenario	Random & Sampling Errors	Conditioning Error	Random, Sampling & Conditioning Errors
2005	lower 90%	1,365	943	912
	expected	1,540	1,540	1,540
	upper 90%	1,716	2,118	2,145
2010	lower 90%	1,507	1,077	1,052
	expected	1,691	1,692	1,691
	upper 90%	1,874	2,320	2,349
2015	lower 90%	1,647	1,179	1,134
	expected	1,841	1,842	1,837
	upper 90%	2,036	2,504	2,547

Note: Forecast intervals for conditioning error only and all three sources of error at the same time are based on Monte Carlo simulations of 1,500, 1,600, and 1,400 iterations for the years 2005, 2010, and 2015, respectively.

A NOTE ON SPECIFICATION ERROR

The previous sections referred to specification error but did not try to quantify it. This is because specification error is by its very nature nonquantifiable. Because the real world is made-up of many complex and inter-related factors that cannot be modeled precisely, it is probably safe to say that in essence most regression models are misspecified. For example, the regression models developed in this chapter almost certainly omit some factors that have an impact on the number of barge tows passing Chadwick Lock. Even more troublesome is the fact that there is no diagnostic test that can establish whether one model specification is "more correct" than another (Kexel, 1988). For example, one cannot compare the R^2 of a model that is linear in relative price and total production to the R^2 of a model that is linear in the logarithms of these variables. Under these circumstances, the focus of model development should be in the specification of models that *adequately* portray these inherently complex relationships, given the set data that is available. Unfortunately, specification of an adequate model is often an innovative/imaginative process of discovery that cannot simply be taught (Kennedy, 1992).

At a minimum, it is important to acknowledge the possibility of model misspecification and to alert decisionmakers of where a particular model may be at fault or where it could use some refinement. There are many statistical techniques that are available to help determine whether any particular regression model is misspecified. For example, the Chow test can be employed to determine whether model coefficients are stable over different time periods. The Box-Cox technique can be used to select the appropriate functional form of a regression model. Model error terms can be tested to see if they are truly random or whether they vary systematically. A full explication of these and the numerous other techniques for diagnosing and improving model misspecification is beyond the scope of this guidebook, and is better left to econometric textbooks. However, it is worth repeating here that all forecasting models have possible shortcomings. One should reveal possible shortcomings, even if a particular forecast model is performing adequately, and particularly, if time and budget constraints limit the model's refinement.

SUMMARY

This chapter defined two approaches to forecasting future movements of barge tows using regression analysis. The first approach used a single independent variable, time, to model the annual number of tows passing Chadwick Lock. Similar to the growth rate approach of Chapter III, this approach openly ignored the effects that actually influence use of the waterway. The second approach defined a regression relationship of two independent variables (crop production and relative price of barge transport), which were hypothesized actually to cause year-to-year changes in the number of tows passing Chadwick Lock.

Four sources of error were identified as inherent sources of uncertainty associated with use of the regression technique. Sampling and random errors were treated in both applications of the regression technique using standard statistical formulae. Unlike the single-variable model that used time as the sole independent variable, conditioning error was present in the multiple regression framework, since forecast values of tows were conditioned on uncertain future values of relative price and total grain production. The combined effects of conditioning, random, and sampling error on the forecast were accounted-for through the use of a Monte Carlo simulation approach. Finally, the single-variable model was considered misspecified, since it overlooked phenomena that are known to affect waterway movements. Despite its improvement in specification, the multiple regression model most likely did not (and could not) account for all of the systematic influences on barge traffic.

VI. THE "TOP-DOWN" APPROACH

INTRODUCTION

The objective of this chapter is to develop a forecast of the number of tows passing Chadwick Lock and to assess its certainty for the years 2005, 2010, and 2015, using a "top-down" approach. The "top-down" approach gets its name from the structure of analyses it entails. Fundamental to this approach is the attempt to specify the main linkages and factors that affect and drive the variable of interest. The *top* often refers to the broad macroeconomic, aggregate supply-demand, phenomena that drive production and output over the long run, while the *down* can relate to the microeconomic and physical relationships that explain short run movements in the forecasted variable. Alternatively, the *top* may represent a logical starting point for identifying and tracing linkages, or interdependencies, all the way *down* to the variable of interest.

Of the four forecasting approaches applied in this manual, the top-down approach is the most ambitious. As shown below, this method relies on Monte Carlo simulation, regression, fitting of probability distributions, and other data analyses. Thus, the top-down approach combines many of the elements described in earlier chapters. If one can learn and apply the methodologies and techniques that are used to analyze risk and uncertainty within the top-down approach, then one will be able to perform risk and uncertainty analysis in most settings.

STOCHASTIC CAUSALITY TREE

What makes up a top-down analysis of the water borne transport of agricultural commodities? Just consider Figure VI-1, where a "causality tree" diagram is presented. The set of branches represents a variety of (stochastic) factors that contribute to the number of barge tows that pass Chadwick Lock, and therefore, to uncertainty in forecasts of future tows. The analytical process described in the figure may be represented by a function:

$$Number\ \ of\ \ Tows\ = f\ (Acres\ Planted,\ Yield,\ ...,\ Modal\ \ Choice,\ etc\ .) \qquad \textbf{(6.1)}$$

where the factors that influence the outcome of the tree are shown as variables.[25] It should be understood that the structure of the tree would be reflected in an algebraic form of (6.1). For

[25]Note in Figure VI-1 that this analysis predominantly traces production, or supply-side, factors. One could extend the tree to identify both domestic and world grain demand factors that influence grain production, as well as the variables that influence these demands. When doing so, one ultimately finds feedback (or an *endogenous* relationship) between supply and demand. The subsequent top-down analysis assumes that grain production accommodates the demand for grain without specifying particular demand factors among the arguements of (6.1).

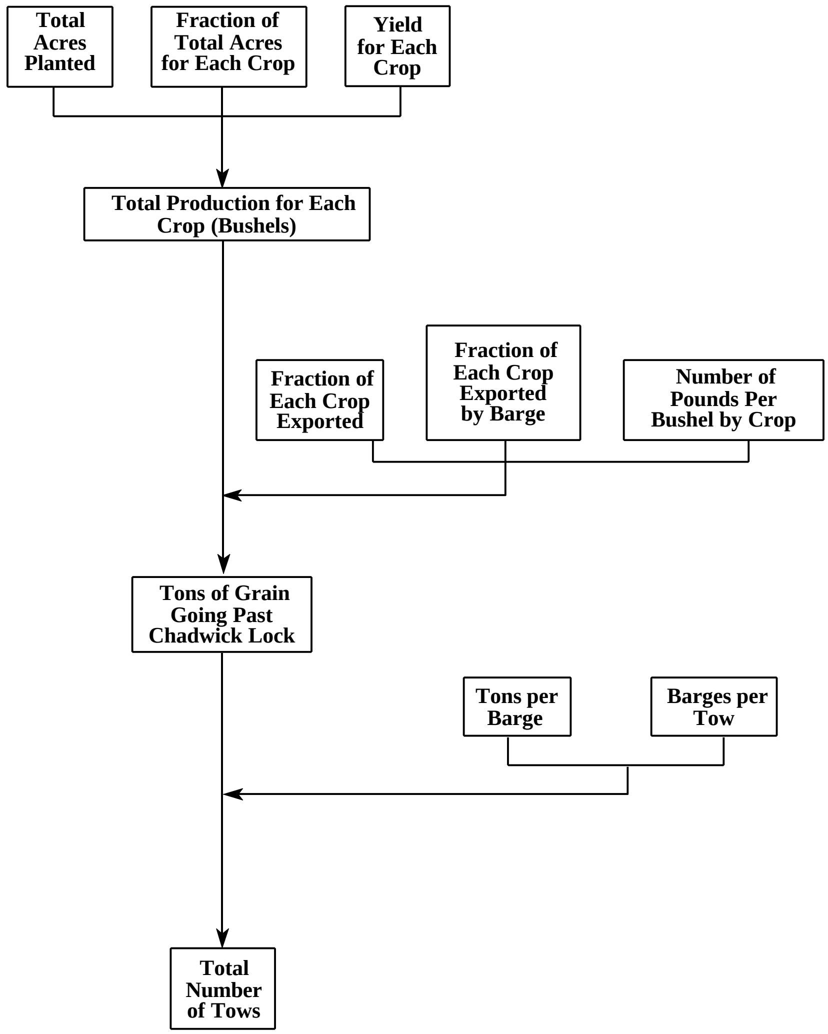

FIGURE VI-1

TREE STRUCTURE OF ANALYSIS IN APPLICATION OF TOP-DOWN APPROACH

instance, if it is possible to represent any one of the variables (like yield) as a function of other new variables, then the variable in the equation would be replaced by this function. In the tree diagram, this would be equivalent to extending the length of a particular branch. More details on the particular choice of the variables and their relations are given later.

How does the causal tree determine the outcome? The causal tree is just a more dynamical representation of Equation (6.1). Fixing the variables at the branches' tips at particular values determines the outcome represented at the base of the tree. If there is uncertainty in the branches of the tree, then there is uncertainty at the base.

MONTE CARLO SIMULATION

Ideally, and knowing exactly the causal relation of (6.1) and the values of its variables, one can easily determine the value of the outcome. In reality, however, the variables of the system can only be predicted with some uncertainty. Therefore, the analysis of such a system should return both the expected value of the outcome and a description of the uncertainty around it. In theory, a full analysis of uncertainty requires an expression of probability. In other words, the values of the variables of Equation (6.1) are not fixed and are more accurately represented by probability distributions. The expected values of these probability distributions take the place of the "certain" values. Therefore, the problem may be stated as follows: given the probability distributions for a set of variables (i.e., the arguments of Equation (6.1)), construct the probability distribution of the function of these variables (i.e., the outcome of Equation (6.1), Number of Tows). This problem is illustrated in Figure VI-2.

A mathematical solution to the problem outlined above is not always possible, and is usually limited to rather simple cases.[26] The method of Monte Carlo simulation presents a solution to the problem.

The Monte Carlo method is a technique of randomly selecting numbers from a given probability distribution that characterizes the underlying data and of obtaining outcomes of the functional relationships among variables. Through repeated sampling from given probability distributions, the technique is able to *simulate* a range of outcomes and closely approximate a probability distribution of the outcome. Each sampling of a simulation is called an *iteration*.

Referring to Figure VI-2, the Monte Carlo method consists of randomly selecting values for each of the variables (total acres, yield, etc.) and calculating the subsequent value of the number of tows. Thus, each of the variables subject to uncertainty must have a defined probability distribution assigned to it.

As the number of iterations grows and the randomly selected values of the variables approximate more closely the assigned distributions, the calculated values of tows reveal more accurately the resulting (conditional) probability distribution. Once the probability distribution on tows is generated, one can derive a statistical description of uncertainty.

Some Technical Issues

The choice of a probability distribution for a variable depends on what one knows about the particular variable. To gain insight into the character of a variable usually requires study of historical data and an understanding of the factors that drive it. The analyst should make every

[26] As pointed out by Pindyk and Rubinfeld (1981) and Kocik et al. (1993a, 1993b), even in more simple cases, mathematical solution is not computationally trivial.

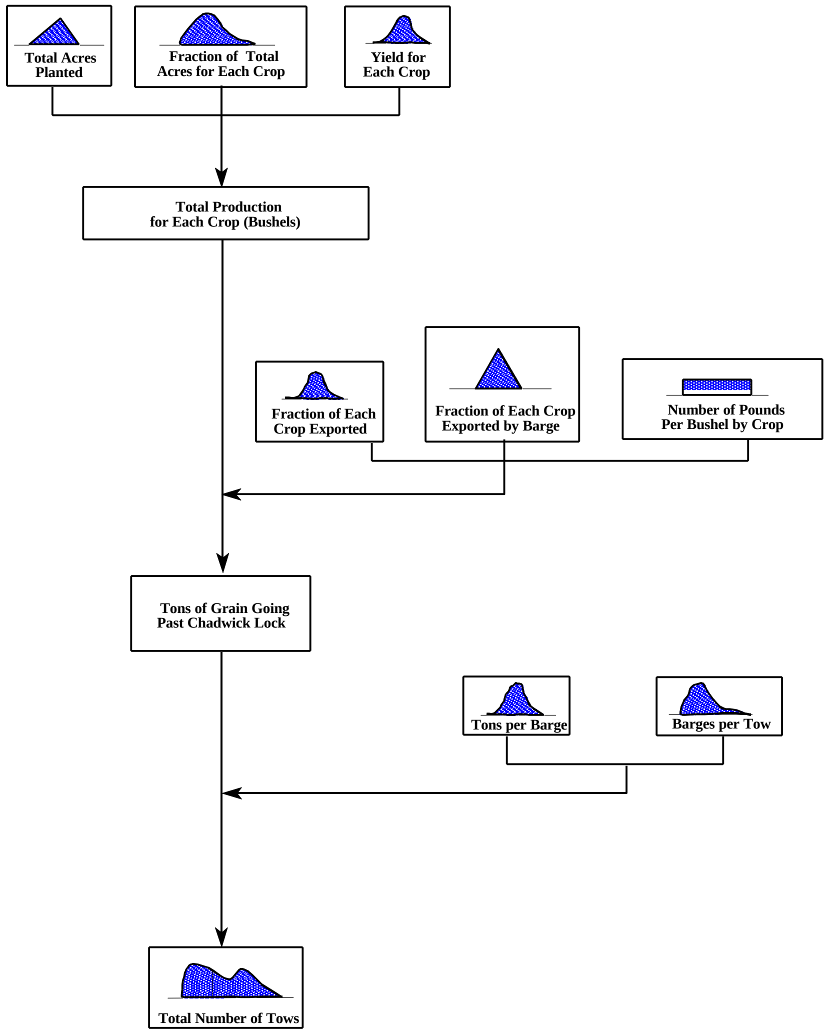

FIGURE VI-2

INTERACTION OF STOCHASTIC DISTRIBUTIONS

attempt to exploit available information about a particular variable in defining its probability distribution. If not much information is available (or if the information is too costly), then the normal distribution may be the best choice, since it requires only two parameters to define its shape (namely mean and variance).

The choice of how many times one should sample from the predefined distributions depends on the degree of accuracy one wants to achieve. The theory of the Monte Carlo method assures that one will eventually *converge* to a certain degree on the underlying probability distributions. A system is said to converge, when an increase in the number of iterations does not significantly contribute to the change in the shape of the distribution that is being simulated. The number of iterations required to reach convergence will vary depending on the complexity of the problem. Another issue related to convergence concerns the small probability event. If a particular event has a low probability of occurrence (say, a 0.001 chance of happening), keep in mind that the model may converge without ever realizing (or sampling) the event. If such an event brings with it high costs and large consequences, then the simulation results should be reviewed to verify that the event occurred in the simulation.

Some of the variables in a simulation model may be correlated. Under these circumstances, the probability distributions of these variables should be described jointly. Often, a correlation matrix generated from statistical software is sufficient to infer a joint probability distribution.

The Monte Carlo simulation technique has been implemented in a number of commercially available computer software packages, which makes an analysis of uncertainty of such complex problems feasible. Because of its compatibility with spreadsheets, the @Risk© software was selected and used for this study. @Risk© provides a range of many different theoretical probability distributions and many user-defined controls for the simulation process. Besides allowing the specification of individual distributions, @Risk© allows one to specify correlations between variables in a simulation model. Furthermore, @Risk© monitors the convergence of simulation outputs and automatically determines the number of iterations that are required.

EXAMPLE APPLICATION

In order to formulate a top-down analysis of the number of tows passing Chadwick Lock, one must envisage the ingredients that make up the components of the tree of Figure VI-1 and VI-2. Especially, as the tree diagram shows, one must determine how much grain will be produced in the Evanstown BEA during any particular period, and how much of this grain will be shipped down the Oak River and its tributaries. The following sections discuss the factors that are expected to influence these variables and develop the equation that will be simulated to prepare the forecast of tows passing Chadwick Lock.

Production

There is ultimately a physical limit to the amount of land in the geographical area of the Evanstown BEA that can be brought into cultivation to produce grain. In the short run, agricultural producers may cultivate marginal lands, retire once farmed lands, reduce or stockpile grain inventories, or bring acreage out of retirement depending on the demand for grains. Over the long-run, larger parcels of noncultivated land may be converted into farmland or vice versa depending on social and economic factors such as trends in urbanization (or sub-urbanization), government programs that subsidize the agricultural sector, and domestic and world demand for agricultural products.

The general *mix* of crops that is planted over time depends on economic conditions such as the relative price received for each crop, and physical factors such as type and quality of soils, and the inertia of reacting to changing demand conditions. The productivity of the cultivated land is measured by yield, which generally differs from crop to crop. Aside from the type and quality of soil, the average number of bushels produced per acre for a particular crop depends on weather conditions from season to season, and on technological advances of seed hybrids and innovations in farming techniques over longer periods of time. Thus, one has the principal short and long run determinants of total grain production—amount of land cultivated, type/mix of crops, and crop yield—which themselves are determined by various short- and long-term factors. In algebraic terms, the amount of production of any particular crop could be defined by the following equation:

$$Production_i = (Total\ Acres\ Planted) \times (Fraction\ of\ Acres\ Planted)_i \times (Yield)_i \qquad \textbf{(6.2)}$$

where each variable in parentheses could be further defined as a function of physical and economic factors and the subscript *I* identifies the type of crop (e.g., corn, wheat, soybeans, other).

Grain production responds to changes in grain demand and is assumed here to fully satisfy this demand. The quantity of domestic grain commodities demanded is dependent on price, which is influenced directly by purchases of raw grain on wholesale markets and indirectly on secondary markets through purchases of retail and other finished products derived from grain. World demand for U.S. grain is primarily made up of sales of raw grain to developing countries and developed countries that do not share the United States' comparative advantage in agriculture.

Barge Shipments

After one has predicted how much grain will be produced, how much will be shipped by barge down a stretch of river? This depends on what amounts of grain needs to go where and on the alternative means through which these amounts can be transported to their appropriate destinations. Typically, a waterway will compete with other modes of transport as a route to transport goods *within* a given region and as a shipping route to destinations *outside* of the region's borders. If the focus of analysis is traffic past a downstream lock or system of locks, one may be interested only in determining the amount of grain "exported" out of the region by barge. The choice of mode of transport, or *modal choice*, depends on a number of factors, such as accessibility, relative unit transport cost, dependability, type of product, and capacity of the given transportation systems.[27] In the agricultural context, both the amount of grains exported out of a region and the amount being transported by barge will likely vary by crop. Ignoring the potential for adjustments to grain inventories, the amount of a particular grain (in tons) passing a downstream lock may also be expressed as an equation:

[27]For a review of the modal choice literature see the IWR report entitled *Transport Mode Selection and Inland Waterborne Commerce* (Harrington and Willett, 1996).

$$
\begin{aligned}
\textit{Tons Passing Lock}_i \quad = \quad & (\textit{Production in bushels})_i \times \\
& (\textit{Fraction of Production Exported})_i \times \\
& (\textit{Fraction of Exports Shipped by Barge})_i \times \\
& (\textit{Pounds per Bushel})_i \div 2000
\end{aligned}
\qquad \textbf{(6.3)}
$$

where, again, the terms on the right-hand side are themselves functions.[28]

Derivation of the Equation for the Top-Down System

Figures VI-1 and VI-2 show that the top-down story does not necessarily end with the modal choice. If one can determine how much of which grain will be transported via the waterway, how does one translate this into a number of barges or barge tows? Thankfully, the answer to this question is not too difficult, that is, if one is willing to forego some precision. One may simply divide the amount of grain by the average amount of grain that fits on a barge to estimate the number of barges. Correspondingly, one may then divide this quantity by the average number of barges per tow to estimate the number of tows.

The principal components of the top-down analysis just described may now be arranged into a single equation that can be used to calculate or predict the number of tows passing Chadwick Lock:

$$
\textit{Number of Tows} = \frac{\displaystyle\sum_{i=1}^{4} \left(\begin{array}{l} (\textit{Total Acres Planted}) \times \\ (\textit{Fraction of Acres Planted})_i \times \\ (\textit{Yield})_i \times \\ (\textit{Fraction of Production Exported})_i \times \\ (\textit{Fraction of Exports Shipped by Barge}) \times \\ (\textit{Pounds per Bushel})_i \div 2{,}000 \end{array} \right)}{(\textit{Number of Tons per Barge}) \times (\textit{Number of Barges per Tow})} \qquad \textbf{(6.4)}
$$

where the numerator represents total tonnage of grain, and where the denominator transforms tonnage into barge tows. This is the equation that specifies the function (6.1). The uncertainty about the values of all of the parameters in this equation needs to be described in terms of probability distributions as in Figure VI-2.

[28] This study assumes that there are 56 pounds per bushel of corn and 60 pounds per bushel of other grains. As the equation suggests, there are 2,000 pounds per short ton.

SPECIFYING THE PROBABILITY DISTRIBUTIONS

The following sections describe how the data inputs to Equation (6.4) were derived and explains the steps that were taken to identify the uncertainty in the inputs of each step of the top-down analysis prior to simulating the system. Appendix A contains the historical data on the input variables under consideration within the top-down system.

Total Acres Planted

Historical data and archives were analyzed to determine the historical variation in land devoted to grain production. Table VI-1 summarizes the historical data. The lowest number of acres employed for grain production in the historical record was approximately 16,600,000 acres. It is considered very unlikely that planted acreage would fall below this amount over the planning horizon, since this would entail an unexpected and steep decline in grain demand and/or dramatic urbanization of once farmed land. The maximum number of acres employed for grain production in the historical record was just under 22,000,000 acres, which represented a period of relatively high exports and no government set-aside programs.

According to many experts, agriculture production is likely to increase steadily over the forecast horizon. Government sponsored acreage set-aside programs will gradually be phased out over the period, which will likely lead grain producers to plant more acres to maintain farm revenues.

TABLE VI-1
TOTAL ACRES PLANTED:
SUMMARY OF HISTORICAL DATA

Mean	19,576,962
Standard Deviation	1,539,632
Minimum	16,590,952
Maximum	21,700,746

Furthermore, recent developments in world trade are favorable to the U.S. grain industry. The GATT and NAFTA trade agreements reduce trade barriers and enhance U.S. producers' competitive advantage in world grain markets. Higher demand for U.S. grain will likely result in more farmland being brought into cultivation.

These expectations of future market conditions be captured using a *triangular* probability distribution for the future values of total acres planted in the Evanstown BEA. The specification of a Triangular Distribution requires three parameters, namely the minimum value, the most likely value, and the maximum value. To reflect changes in government agricultural policy and international trade agreements, the most likely values for total acres planted were assumed to approach the maximum historical value over the forecast period.

This anticipated trend is shown in the data of Table VI-2, which reports the minimum, the most likely, and the maximum values of the respective assumed Triangular Distributions for the forecast years 2005, 2010, and 2015. Meanwhile, Figure VI-3 illustrates the shape of the assumed triangular distribution of acres planted, and shows how the shape is expected to change over time.

TABLE VI-2
ASSUMED TRIANGULAR DISTRIBUTIONS
FOR TOTAL ACRES PLANTED (acres)

Forecast Year	Minimum Value	Most Likely Value	Maximum Value
2005	16,600,000	20,500,000	22,000,000
2010	16,600,000	21,000,000	22,000,000
2015	16,600,000	21,500,000	22,000,000

Fraction of Total Acres Planted by Crop

The total number of acres planted is divided among corn, wheat, soybeans, and other grains. There is no clear consensus on what the future relative shares of these crops will be. Therefore, the historical data on the relative mix of four grain commodities was analyzed in order to assign appropriate probability distributions from which to randomly select future crop shares within the Monte Carlo analysis.

Summary statistics for the historical data are reported in Table VI-3. Historically, corn has tended to dominate the other crops in terms of relative share of acres planted, followed in order by soybeans, wheat, and other grains. Except for the other grains category, there has been relatively low variation in the shares devoted to each crop.

Correlation analysis was undertaken to study inter-relationships among the relative shares of each crop and with their lagged values and time. Table VI-4 shows the results of the correlation analysis. Statistically significant correlations are found to exist between the percentages of acres devoted to soybeans and to corn, as well as between the percentages of acres

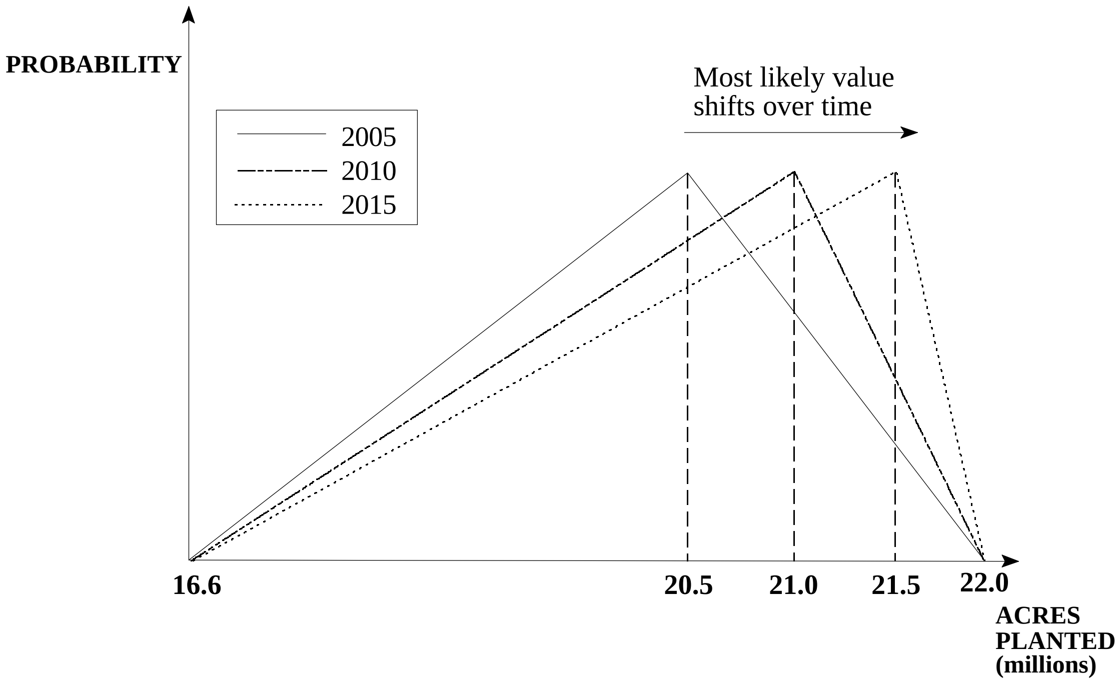

FIGURE VI-3

ILLUSTRATION OF USE OF TRIANGULAR DISTRIBUTION

TABLE VI-3
FRACTION OF TOTAL ACRES PLANTED: HISTORICAL AVERAGES BY CROP

Crop	Mean Fraction of Total Acres	Standard Deviation	Minimum	Maximum
Corn	0.5125	0.0202	0.4395	0.5366
Wheat	0.0675	0.0127	0.0384	0.1005
Soybeans	0.3580	0.0357	0.2834	0.4258
Other Grains	0.0620	0.0268	0.0255	0.1256

TABLE VI-4
CORRELATION MATRIX OF FRACTION OF TOTAL ACRES BY CROP
Pearson Correlation Coefficients
(Prob>|R| under HO:Rho=0)

	% CORN	% SOYBEAN	% WHEAT	% OTHER	YEAR	LAG % CORN	LAG % WHEAT	LAG % SOYBEAN	LAG % OTHER
% CORN	1.0000 0.0000	*-0.6602* *0.0002*	0.0530 0.7971	0.0995 0.6287	-0.1370 0.5045	-0.0195 0.9262	0.1756 0.4012	-0.1264 0.5471	0.0941 0.6545
% SOYBEAN		1.0000 0.0000	-0.2182 0.2843	*-0.7294* *0.0001*	*0.7671* *0.0001*	-0.0691 0.7428	-0.1568 0.4541	*0.6232* *0.0009*	*-0.6827* *0.0002*
% WHEAT			1.0000 0.0000	-0.2222 0.2753	0.0550 0.7896	-0.0113 0.9574	*0.5409* *0.0052*	-0.1165 0.5792	-0.1039 0.6212
% OTHER				1.0000 0.0000	*-0.9428* *0.0001*	0.1230 0.5581	-0.1783 0.3938	*-0.7475* *0.0001*	*0.9671* *0.0001*
YEAR					1.0000 0.0000	-0.0907 0.6665	0.0951 0.6513	*0.7378* *0.0001*	*-0.9388* *0.0001*
LAG % CORN						1.0000 0.0000	0.0375 0.8588	*-0.6527* *0.0004*	0.0586 0.7807
LAG % WHEAT							1.0000 0.0000	-0.1962 0.3474	-0.2623 0.2053
LAG % SOYBEAN								1.0000 0.0000	*-0.7030* *0.0001*
LAG % OTHER									1.0000 0.0000

Note: Correlations in *italics* are significant at the 0.10 level or higher.

devoted to soybeans and to other grains. In general, as a higher proportion of total acres available is devoted to soybean production, less is devoted to the production of corn and “other” grains.

The correlation analysis also indicates that the shares of total acres devoted to wheat, soybeans, and other grains, are correlated with the relative shares of these crops in past periods, as represented by the one-year lags of these variables. This might suggest a degree of inertia in changing the mix of crops planted, however it is unclear from the data whether this pattern is due to technical difficulties in changing crop mix from year to year or slowly changing trends in the economic factors that influence crop mix, or both. Because of these unknowns, this information was not built into the simulation.

Finally, the shares of total acres devoted to soybeans and other grains show a significant time trend. The fraction of total acres devoted to soybeans has been increasing over time, apparently at the expense of the number of acres devoted to the other grains category. Similar to the lagged relationships, it is not clear from the data what is behind this trend. One could hypothesize a number of functional relationships between the fraction of acreage devoted to each crop and variables such as relative price received for each crop, local demand, export demand, and others that operate in time. However, the cost of collecting data on these variables was prohibitive and prevented the development and specification of predictive regression models within the simulation.

The historical data tend to indicate fraction of acres devoted to each crop is fairly concentrated around the historical means. This is evident in the four panels of Figure VI-4, which plot the frequency histograms of the historical shares of total acres devoted to each crop. The smooth curves in the diagram represent theoretical probability distributions fitted to the historical data using the BestFit© computer software program. As shown, the historical shares of corn are best represented by a Weibull probability distribution that is skewed to the left, but concentrated around the value of 0.52. The historical values for the relative share of wheat and soybeans are each represented Normal distributions defined by the historical mean and standard deviation. Finally, the historical values of the share of other grains is fitted best with a Log-normal distribution, which is skewed to the right and concentrated around the historical mean value of 0.062. Because of the relatively small amount of historical variation in the fraction of total acres planted in each crop, and in absence of other predictive relationships, these fitted theoretical distributions were assigned to portray the future variation in the fraction of total acres devoted to each crop. The @Risk© simulation software allowed direct entry of the theoretical distributions based on the parameters estimated by the BestFit© program, which define the shapes of the distributions.[29] Further, in order to account for covariance among the fraction of total acres devoted to each crop, the statistically significant correlations between the individual fractions of total acres of each crop were input into the Monte Carlo simulation using @Risk©'s *Correlate* option. This permits sampling from one theoretical distribution to affect the sampling of the other distributions during the simulation.[30]

Crop Yield

Crop yield is assumed to be a function of type of crop, weather, and technological development. Sufficient data was available to model the yield of each crop using multiple regression. The following linear equation was estimated for each crop:

---

[29] For example, for the theoretical probability distribution for the fraction of total acres devoted to corn, the term "=RiskWeibull(36.2, 0.52)" is entered into the @Risk© spreadsheet. The two numeric parameters, 36.2 and 0.52 define the shape of the fitted Weibull distribution as shown in Figure VI-4.

[30] Within the simulation spreadsheet, the sampled values for the fraction of total acres planted for each crop are adjusted (i.e., scaled) so that they add to 1.0 when summed across crops.

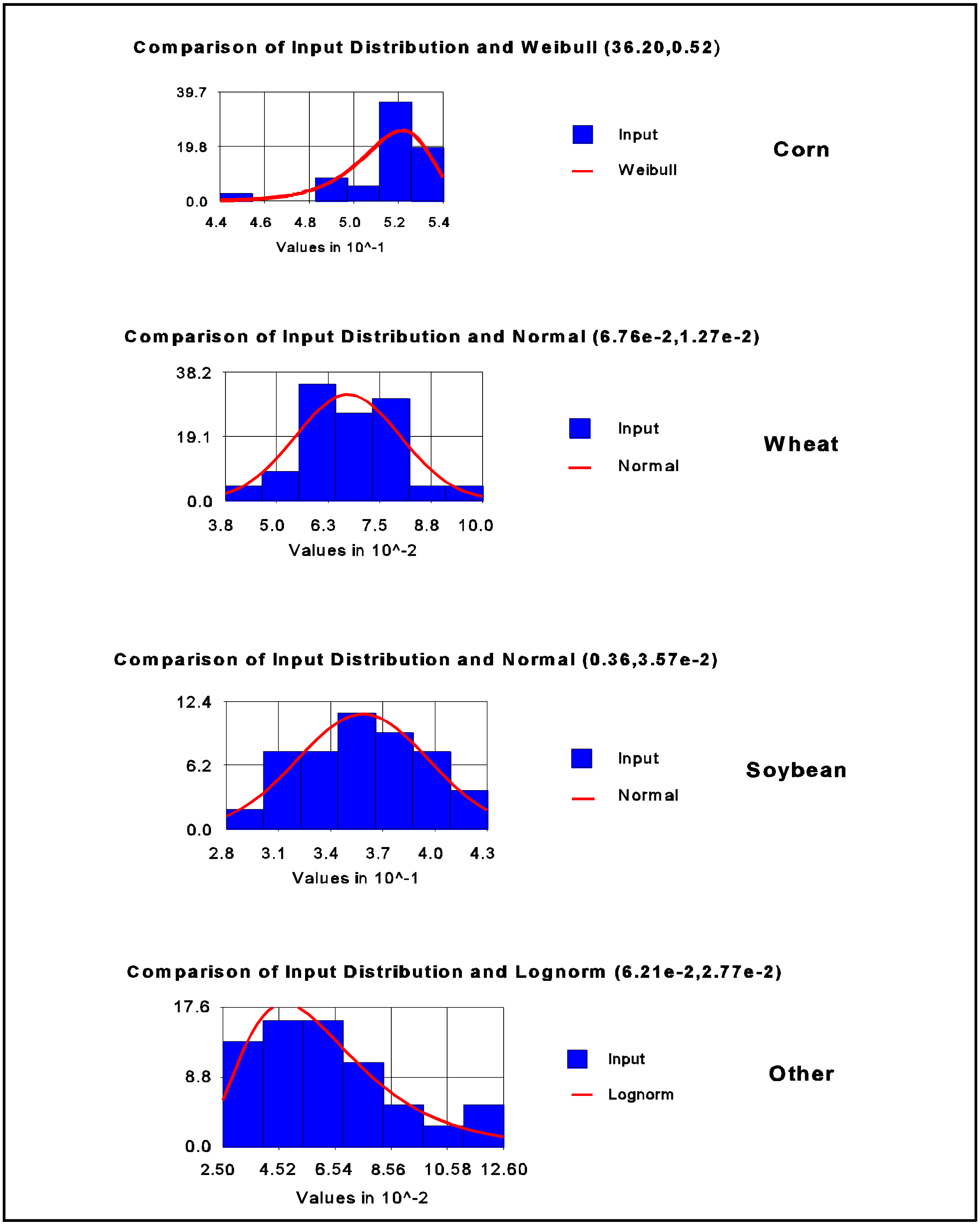

FIGURE VI-4

FREQUENCY HISTOGRAMS OF HISTORICAL DATA AND THEORETICAL PROBABILITY DISTRIBUTIONS: FRACTION OF TOTAL ACRES PLANTED BY CROP

$$Yield = \alpha + \beta_1 CDD + \beta_2 Time + \varepsilon \qquad \textbf{(6.5)}$$

where:

Yield	=	number of bushels harvested per acre
α	=	unknown model intercept term
CDD	=	cooling degree days (annual total)
Time	=	defined as (year - 1995)
β_1,β_2	=	unknown slope parameters
ε	=	random error term

The time variable is specified to serve as a proxy for technological change, and may be expected to retain a positive coefficient estimate. The weather variable, annual number of cooling degree days, is specified to capture the effects of evapotranspiration–the more cooling degree days, the higher the evapotranspiration, and the higher the yield.[31]

The parameters of equation (6.5) were estimated for each crop using the SAS© regression procedure. The results are reported in Tables VI-5 (corn), VI-6 (wheat), VI-7 (soybeans), and VI-8 (other). As expected, crop yields are found to be positively related to both time and cooling degree days. The estimated coefficients for the intercept term imply that corn yields are generally the highest, followed by other grains, wheat, and soybeans, respectively. The model for soybean yields has the best fit among the models as suggested by a comparison of the values of R^2. Generally, however, most of the variation in crop yields is left unexplained.

As explained in Chapter V, the estimated regression parameters shown in Tables VI-5 through VI-8 can be used to predict future values for yield by crop by substituting assumed values for explanatory variables for each forecast year into the relationship. As also explained in Chapter V, the predictions from the regression relationship will contain sampling, random, and conditioning error.

The regression equation becomes an algebraic part of Equation (6.4) in the simulation process. It is standard procedure to describe the uncertainty of the regression equation in terms of the normal distribution. To account for sampling error, the estimated parameters are assumed to vary normally around their expected values (i.e., the values shown in the parameter estimates column) according to the their standard errors (i.e., the values shown in the third column of the tables). Recall also from Chapter V, that estimates of forecast error should incorporate covariance among the model parameters. The matrices provided at the bottom of the tables of

[31]Cooling degrees are the number of degrees (F) by which the average temperature for the day exceeds 65 degrees (F). For example, if the average temperature is 70, then there are 5 cooling degree days recorded for that day. Days in which the average daily temperature is less than 65 receive a value of zero for CDD. To correspond to the time step of the regression analysis, cooling degree days are summed by year.

TABLE VI-5
REGRESSION RESULTS FOR CORN YIELD

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Prob > |T|
INTERCEPT	78.090389	34.31081017	2.276	0.0325
TIME	1.008272	0.43703851	2.307	0.0304
CDD	0.029402	0.02110888	1.393	0.1770

Correlation of Estimates

	INTERCEPT	TIME	CDD
INTERCEPT	1.000	0.2508	-0.9828
TIME	0.2508	1.0000	-0.0946
CDD	-0.9828	-0.0946	1.0000

N = 26
Adj. R^2 = 0.1920
F-value = 3.971
Prob > F = 0.0330
Root MSE = 16.63853

TABLE VI-6
REGRESSION RESULTS FOR WHEAT YIELD

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Prob > |T|
INTERCEPT	29.634957	11.56238334	2.563	0.0174
TIME	0.310001	0.14727740	2.105	0.0464
CDD	0.008793	0.00711347	1.236	0.2289

Correlation of Estimates

	INTERCEPT	TIME	CDD
INTERCEPT	1.000	0.2508	-0.9828
TIME	0.2508	1.0000	-0.0946
CDD	-0.9828	-0.0946	1.0000

N = 26
Adj. R^2 = 0.1528
F-value = 3.255
Prob > F = 0.0569
Root MSE = 5.60701

TABLE VI-7
REGRESSION RESULTS FOR SOYBEAN YIELD

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Prob > |T|
INTERCEPT	29.34157	7.06597174	4.153	0.0004
TIME	0.400340	0.09000376	4.448	0.0002
CDD	0.007138	0.00434716	1.642	0.1142

Correlation of Estimates

	INTERCEPT	TIME	CDD
INTERCEPT	1.000	0.2508	-0.9828
TIME	0.2508	1.0000	-0.0946
CDD	-0.9828	-0.0946	1.0000

N = 26
Adj. $R^2 = 0.4690$
F-value = 12.039
Prob > F = 0.0003
Root MSE = 3.42654

TABLE VI-8
REGRESSION RESULTS FOR OTHER YIELD

Variable	Parameter Estimate	Standard Error	T for HO: Parameter = 0	Prob > |T|
INTERCEPT	34.958448	13.90790040	2.514	0.0194
TIME	0.356282	0.17715373	2.011	0.0562
CDD	0.017284	0.00855649	2.020	0.0552

Correlation of Estimates

	INTERCEPT	TIME	CDD
INTERCEPT	1.000	0.2508	-0.9828
TIME	0.2508	1.0000	-0.0946
CDD	-0.9828	-0.0946	1.0000

N = 26
Adj. $R^2 = 0.2181$
F-value = 4.487
Prob > F = 0.0226
Root MSE = 6.74443

regression output reflect the covariance among the model parameters, in terms of correlation coefficients.[32] These correlations were incorporated into @Risk© to account for the covariance among the parameter estimates.[33]

To account for random error, a normally distributed error term is added to the prediction of the equation. The error term is assumed to have a mean of zero and standard deviation equal to the standard error of the regression equation (i.e., the RSME or "root mean square error" term provided in the regression output).[34] The power of the Monte Carlo method also allows one to incorporate the conditioning error introduced by the variable cooling degree days. Similar to its treatment in Chapter V, the future annual number of cooling degree days is assumed to be normally distributed around its long-term annual average.

Fraction of Production Exported

As mentioned previously, the amount of grain that is produced in any particular year is assumed to be either consumed within the Evanstown BEA, or exported out of the region to other BEA's or to foreign countries via the international export facilities at Cajun City. Thus, for simplicity, it is assumed that there is no grain stored as inventory. For each crop, the amount of grain that is exported is calculated as a fraction (or percent) of the total amount of that grain produced.

Local experts expect the *amount* of exports to rise over the forecast period. However, there is no clear consensus on whether exports will assume a larger *proportion* of production. Agricultural forecasters expect the proportion of each grain going to export will at least maintain past levels. Wheat is expected to maintain its position as the most widely exported grain, followed by corn and by soybeans. This is brought to light by the summary of historical data in Table VI-9. The correlation matrix of Table VI-10 further indicates that, historically, the higher the fraction of production of one grain that is exported, the higher is the fraction of other grains that is exported. All of the between-crop correlations are statistically significant at the 0.10 level or higher. Interestingly, only the other grains variable shows a significant correlation with time.

---

[32] Remember from Chapter III the formula for covariance. Algebraically, this formula may be rearranged to calculate a correlation coefficient. On request, SAS© routines allow generation of either the covariance matrix or the correlation matrix.

[33] Since the data for the independent variables are the same for each regression model, this causes the correlations at the bottom of Tables VI-5 through VI-8 to be identical.

[34] Since the yield variables share common determinants, cooling degree days and time, one might suggest that the error terms among the four regression equations are correlated. In other words, an over prediction of corn yield may be associated with an over-prediction of wheat yield. More sophisticated regression routines would be required to detect and account for such correlations if they exist (for example, the Seemingly Unrelated Regression method). For this analysis it is assumed that the errors of the individual models are uncorrelated with one another.

TABLE VI-9
FRACTION OF PRODUCTION EXPORTED: HISTORICAL AVERAGES BY CROP

Crop	Mean Fraction of Production Exported	Standard Deviation	Minimum	Maximum
Corn	0.5958	0.0474	0.4800	0.6800
Wheat	0.7093	0.0607	0.5560	0.8160
Soybeans	0.3446	0.0400	0.2250	0.4072
Other Grains	0.2143	0.0449	0.1200	0.2850

TABLE VI-10
CORRELATION MATRIX OF FRACTION OF PRODUCTION EXPORTED BY CROP
Pearson correlation Coefficients
(Prob>|R| under HO:Rho=0)

	% CORN	% SOYBEAN	% WHEAT	% OTHER	YEAR
% CORN	1.0000	*0.4253*	*0.9283*	*0.4185*	0.2321
	0.0000	*0.0303*	*0.0001*	*0.0334*	0.2539
% SOYBEAN		1.0000	*0.4868*	*0.3835*	0.2669
		0.0000	*0.0117*	*0.0531*	0.1875
% WHEAT			1.0000	*0.3465*	0.3187
			0.0000	*0.0829*	0.1125
% OTHER				1.0000	*0.5465*
				0.0000	*0.0039*
YEAR					1.0000
					0.0000

Note: Correlations in *italics* are significant at the 0.10 level or higher.

The BestFit© program was used to graphically analyze the historical data on the fraction of each grain shipped to export. The four panels of Figure VI-5 show the historical frequency distributions of the variables along with the theoretical probability distributions fitted by the software program. Among the theoretical probability distributions available in the BestFit©, the fractions of corn and wheat production exported are each estimated to fit most closely to a Normal distribution, even though the fit is far from perfect. The fractions of soybean and other grain production shipped to export are both estimated to follow the Weibull distribution. The theoretical distribution for soybeans is skewed to the left and concentrated around the value of 0.36. Meanwhile, the theoretical distribution for the other grains category has a near-normal shape that is concentrated around the value of 0.08. These theoretical probability distributions were assigned within the Monte Carlo simulation to account for uncertainty in future values of the fraction of production exported variables, incorporating the between-crop correlations shown in Table VI-10.

Fraction of Exports Shipped by Barge (Modal Choice)

The summary statistics of Table VI-11 show that barge transport generally has been the grain export mode of choice in the Evanstown region. On average over the last quarter of a century, the Oak River and its tributaries have moved about 87 percent of exported wheat, 75 percent of corn, and 69 percent of soybean exports to the ports of Cajun City. The standard deviations suggest that these fractions have not varied considerably around their averages. This inference is borne out by the small ranges between the historical minima and maxima. Table VI-12 indicates significant positive historical correlations between the fraction of exports shipped by barge among the crops. A correlation is also found to exist between barge shipments of corn, soybeans, and other crops with time.

TABLE VI-11
FRACTION OF EXPORTS SHIPPED BY BARGE : HISTORICAL AVERAGES BY CROP

Crop	Mean Fraction of Production Exported	Standard Deviation	Minimum	Maximum
Corn	0.7448	0.0493	0.6330	0.8310
Wheat	0.8714	0.0444	0.7870	0.9500
Soybeans	0.6878	0.0311	0.6200	0.7320
Other Grains	0.4592	0.0649	0.3500	0.6000

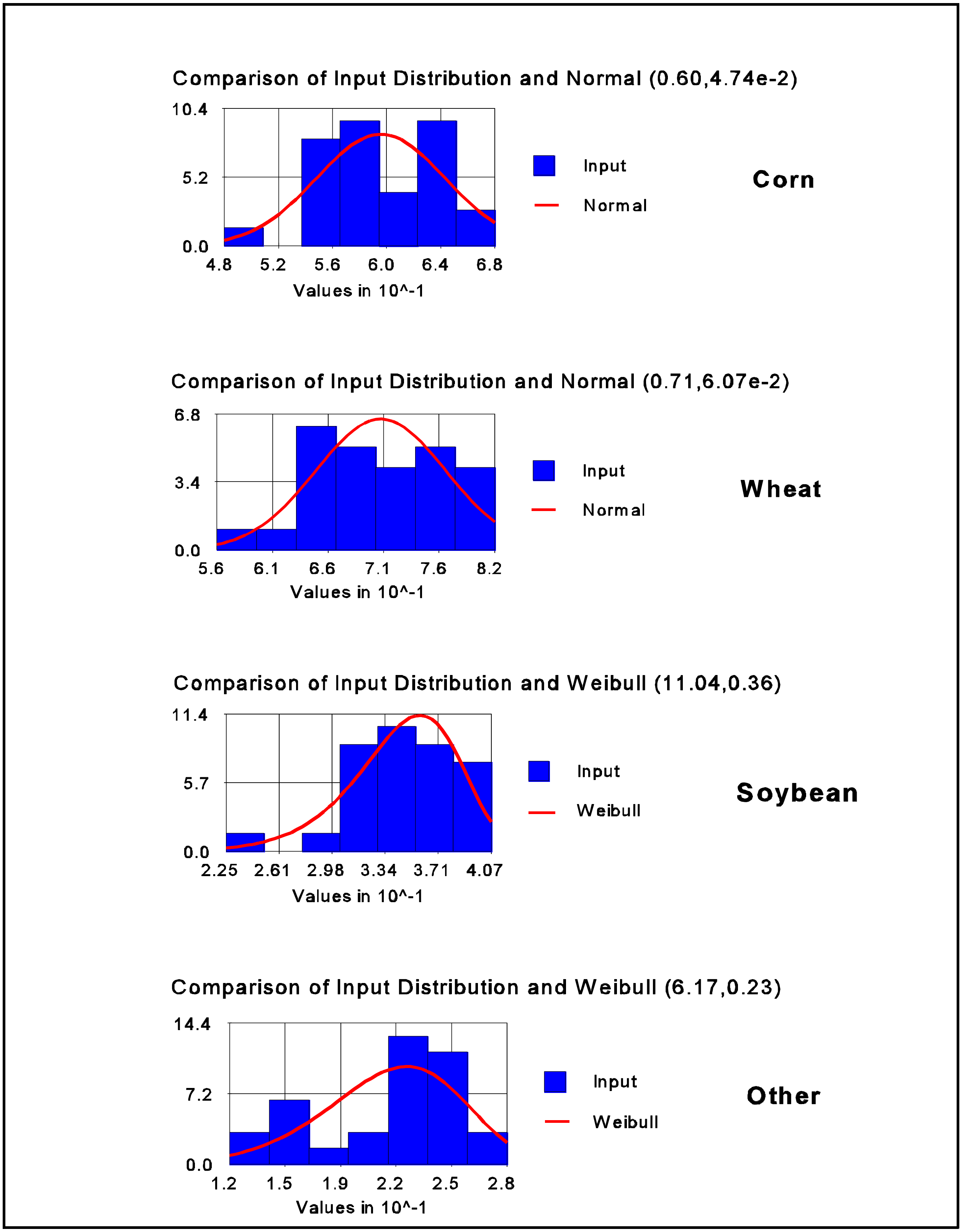

FIGURE VI-5

FREQUENCY HISTOGRAMS OF HISTORICAL DATA AND THEORETICAL PROBABILITY DISTRIBUTIONS: FRACTION OF PRODUCTION EXPORTED

TABLE VI-12
CORRELATION MATRIX OF FRACTION OF EXPORTS SHIPPED BY BARGE
Pearson correlation Coefficients
(Prob>|R| under HO:Rho=0)

	% CORN	% SOYBEAN	% WHEAT	% OTHER	YEAR
% CORN	1.0000 0.0000	0.0088 0.9661	*0.4422* *0.0237*	-0.0404 0.8447	*-0.4082* *0.0385*
% SOYBEAN		1.0000 0.0000	*0.5984* *0.0012*	0.2480 0.2219	*0.6183* *0.0008*
% WHEAT			1.0000 0.0000	0.0513 0.8034	0.1891 0.3549
% OTHER				1.0000 0.0000	*0.3979* *0.0441*
YEAR					1.0000 0.0000

Note: Correlations in *italics* are significant at the 0.10 level or higher.

Shippers in the Evanstown region have a choice to ship grain products for export either by barge or by train (or by both). There are many opinions on what determines the choice of transport mode, and many functional relationships have been hypothesized. In aggregate, this modal choice may be a function of such variables as the type of grain being transported, the ratio of the average cost of barge shipments to the average cost of rail shipments (i.e., relative price), average delay rates, and others. Unfortunately, there is no consensus on what variables best represent and predict transport mode selection. Therefore, it is very difficult to forecast the conditions that will prevail within the transportation market. Any model of modal choice as a simple function of time would be far too misspecified within this framework, and more sophisticated functions would likely be too costly for this exercise, due to the extensive nature of primary data collection.

To account for the potential variation in future values of the modal choice variables within this context, the fraction of exports of each crop shipped by barge was assigned to follow a continuous *uniform* probability distribution. A uniform distribution is defined by only two values, namely a maximum and a minimum. Each of the values of a uniform distribution (including the extreme values and all values in between) have an equal chance of occurring. Using the historical maxima and minima reported in Table VI-11, the four panels of Figure VI-6 illustrate the theoretical uniform distributions that were assigned to the fraction of exports of each crop

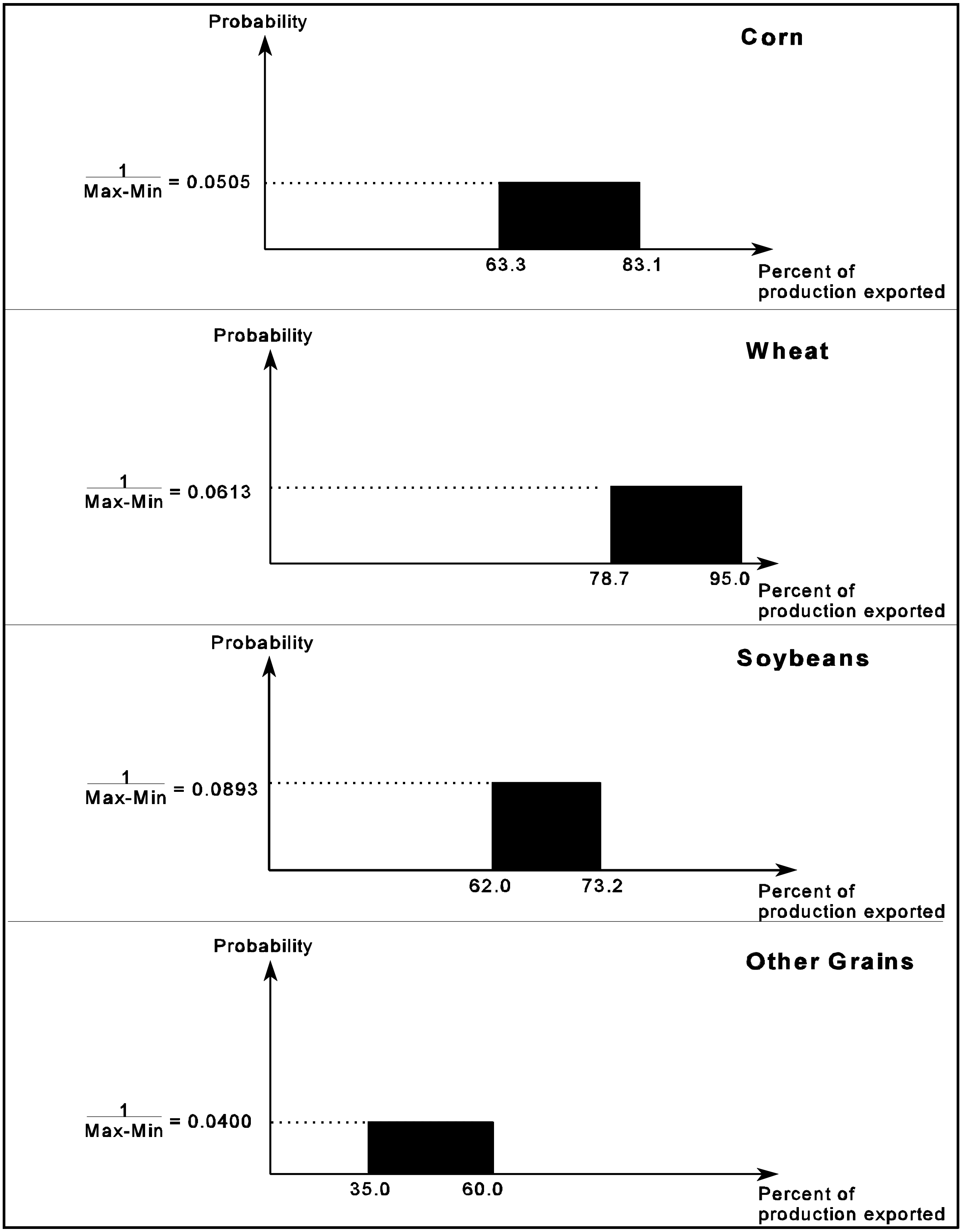

FIGURE VI-6

ASSIGNED UNIFORM DISTRIBUTIONS FOR FRACTION OF PRODUCTION EXPORTED BY CROP

shipped by barge.[35] These probability distributions were input into the @Risk© simulation spreadsheet, and, through the use of @Risk©'s correlate function, incorporate the statistically significant between-crop correlations shown in Table VI-12.

Tons per Barge and Barges per Tow

According to Equation (6.4), the variables (a) number of tons per barge and (b) barges per tow, translate the amount of grain moving down the Oak River into the number of barge tows moving past Chadwick Lock. Long-term records indicate that an average of 1,520 tons of grain is loaded on an average barge and that an average of 10 barges is attached to the average tow passing through the Chadwick Lock. Since these are averages, one knows that variation and uncertainty does exist in these variables. However, in this analysis, the resulting number of tows required to transport the tonnage of commodities is large enough to make the contribution of the variance of these variables negligible. Therefore, the analysis will fix the value of the denominator of Equation (6.4) and not explore uncertainty in these variables.

Defining Correlations Among other Input Distributions

As described above, many statistically significant correlations were found to exist *between crops* for the specific input variables. An additional analysis of the historical data set was undertaken in order to determine whether significant correlations exist *between the input variables* for each crop. For example, statistical tests were performed to determine whether the fraction of acres devoted to corn is correlated with the fraction of corn production exported and with the fraction of corn exports shipped by barge. Table VI-13 shows a correlation matrix of the historical data of the input variables for which probability distributions have been defined. The following variable names are used in the table:

FACORN	=	fraction of total acres devoted to corn production
FAWHET	=	fraction of total acres devoted to wheat production
FASOY	=	fraction of total acres devoted to soybean production
FAOTH	=	fraction of total acres devoted to production of other grains
FECORN	=	fraction of corn production shipped to export

[35] One should not confuse the concept of *probability* with that of *probability density*. As an example of how the continuous uniform distribution is used to calculate probability, consider the first panel of Figure VI-6. Since the distribution is continuous, there are an infinite number of points between the minimum of 63.3 percent and the maximum of 83.1 percent. For any continuous probability distribution, the probability that a *particular point* will occur is zero. However, the probability that the percent of corn production exported falls within an *interval* is definable. For example, the probability that percent of production exported is between 64.3 percent and 63.3 percent is defined by the area under the straight line, which, in the case of a 1 percentage unit change is 0.0505, as labeled on the vertical axis. The formula for the *cumulative* uniform distribution is (x - min)/(min - max), and can be used to calculate the probability of any interval given the extremes of the particular uniform density function.

TABLE VI-13

CORRELATION MATRIX OF TOP-DOWN INPUT VARIABLES
Pearson Correlation Coefficients (Prob > |R| under HO: Kho = 0/N=26)

	PCORN	PWHET	PSOY	POTH	PECORN	PEWHET	PESOY	PEOTH	PBCORN	PBWHET	PBSOY	PBOTH	ATOT	CDD
PCORN	1.0000	0.0530	***-0.6602***	0.0995	0.1553	0.1332	-0.0634	-0.2279	0.2142	0.2864	0.0858	*-0.3594*	*0.5024*	0.3251
	0.0000	0.7971	***0.0002***	0.6287	0.4487	0.5165	0.7584	0.2628	0.2934	0.1560	0.6768	*0.0713*	*0.0089*	0.1051
PWHET		1.0000	-0.2182	-0.2222	0.3236	0.2912	-0.2029	0.1043	-0.0733	-0.2743	-0.2672	0.0434	*0.4572*	0.1934
		0.0000	0.2843	0.2753	0.1068	0.1490	0.3202	0.6120	0.7221	0.1750	0.1870	0.8331	*0.0189*	0.3439
PSOY			1.0000	***-0.7294***	0.0623	0.1296	0.2726	*0.5016*	*-0.3793*	-0.0358	***0.4232***	0.4950	-0.0840	-0.1351
			0.0000	***0.0001***	0.7626	0.5281	0.1779	*0.0090*	*0.0560*	0.8622	***0.0312***	0.0101	0.6832	0.5107
POTH				1.0000	*-0.3527*	*-0.4102*	-0.2189	***-0.5447***	*0.3776*	-0.0386	*-0.5013*	***-0.4080***	*-0.4828*	-0.1567
				0.0000	*0.0772*	*0.0374*	0.2826	***0.0040***	*0.0572*	0.8514	*0.0091*	***0.0385***	*0.0125*	0.4446
PECORN					1.0000	***0.9283***	***0.4253***	***0.4185***	-0.2435	-0.1904	0.1447	-0.0933	*0.5810*	-0.0270
					0.0000	***0.0001***	***0.0303***	***0.0334***	0.2306	0.3516	0.4807	0.6502	*0.0019*	0.8959
PEWHET						1.0000	***0.4868***	***0.3465***	-0.2298	-0.1469	0.1888	-0.0876	*0.5525*	0.0185
						0.0000	***0.0117***	***0.0829***	0.2588	0.4740	0.3556	0.6705	*0.0034*	0.9287
PESOY							1.0000	***0.3835***	0.0058	0.0653	***0.4740***	0.1369	0.1185	0.0110
							0.0000	***0.0531***	0.9777	0.7512	***0.0144***	0.5049	0.5644	0.9574
PEOTH								1.0000	-0.3012	0.0876	*0.4704*	0.1602	0.1915	-0.1591
								0.0000	0.1349	0.6705	*0.0153*	0.4344	0.3486	0.4376
PBCORN									1.0000	***0.4422***	0.0088	-0.0404	-0.0197	0.0408
									0.0000	***0.0237***	0.9661	0.8447	0.9240	0.8433
PBWHET										1.0000	***0.5984***	0.0513	-0.1578	-0.2597
										0.0000	***0.0012***	0.8034	0.4415	0.2002
PBSOY											1.0000	0.2480	-0.0148	-0.0777
											0.0000	0.2219	0.9430	0.7058
PBOTH												1	-0.053	-0.081
												0	0.7985	0.6952
ATOT													1	*0.44661*
													0	*0.0222*
CDD														1
														0

Note: Correlation in *italics* are significant at the 0.10 level or higher. Correlations in **bold** were incorporated into the Monte Carlo simulations.

FEWHET	=	fraction of wheat production shipped to export
FESOY	=	fraction of soybean production shipped to export
FEOTH	=	fraction of other grains production shipped to export
FBCORN	=	fraction of corn exports shipped by barge
FBWHET	=	fraction of wheat exports shipped by barge
FBSOY	=	fraction of soybean exports shipped by barge
FBOTH	=	fraction of other grains exports shipped by barge
ACTOT	=	acres planted
CDD	=	cooling degree days

There are many statistically significant correlations identified in the matrix, some of which repeat the contents of the other correlation matrices presented earlier in this chapter. Some of the results seem intuitive, while others defy direct interpretation. The fraction of total acres devoted to soybeans is positively correlated with the fraction of soybean exports that are shipped by barge. For this finding, one may suggest that as relatively more land and production is associated with soybeans in response to higher export demand, producers must rely on more barge transport, everything else held constant, to get the grain to purchasers. This would appear to be substantiated by the significant and positive relationship between the fraction of soybean production exported (FESOY) and the fraction of exports shipped by barge (FBSOY). On the other hand, the table indicates an inverse relationship between the fraction of total acres devoted to other grains (FAOTH) with both the fraction of other grains exported (FEOTH) and the fraction of these exports shipped by barge (FBOTH). One may hypothesize that increases in production of other grains normally stem from increases in demand from nearby domestic sources, which are less reliant on the waterway for delivery of goods.

Other statistically significant pair-wise (or *bivariate*) relationships are harder to describe, specifically those concerning correlations across both variables and crop types. For example, note the correlations between FAOTH and FBSOY and between FACORN and FBOTH. Intuitively, each of these pairs of variables should not be directly related. It is safe to say that such correlations are brought about through shared correlations with other variables listed in the table or variables not incorporated within the model.[36,37] Other elusive cases involve the total acres planted variable (ACTOT). The correlations indicate that as more acres are planted, a higher proportion of these acres tend to be devoted to corn and to wheat, with a lower proportion going to other grains. The variation in these variables shares common causes, such as changes in the relative prices received for the crops and the existence and degree of farm subsidies and set-aside programs. Thus, the decisions on the amount of acres to plant and the mix of crops to plant are made concurrently. As mentioned earlier, there are expectations that the total amount of acres planted will increase over the forecast

[36]For example, by referring to the correlation matrix the reader may trace the relationship between FAOTH and FBSOY to the negative correlation between FAOTH and FASOY. Similarly, the correlation between FACORN and FBOTH can be traced to the correlations between FACORN and FASOY, FASOY and FAOTH, and FAOTH and FBOTH.

[37] Of course, this says nothing about what actually causes correlation between the variables, which represents a real limitation of bivariate analysis. If more data were available, multiple regression analysis would allow for direct specification and inference with regard to effects of the causal factors discussed within the chapter. Conceptually, all of the variables within the top-down system of this chapter could be specified as estimable functions of causal factors, similar to the treatment of crop yields. The top-down analysis could then be represented by a system of regression equations, the predictions from which would interact to produce the forecast distribution on tows.

period, due to the phase-out of acreage set-aside programs and the effects of world trade agreements. On the other hand, there is no consensus on what to expect in terms of crop mix, due to uncertainties in the make-up of domestic and world demand for grain. Thus, there is ample reason to overlook the historical correlation among these variables. Finally, a positive correlation exists between total acres planted and the annual number of cooling degree days (CDD). Since cooling degree days accumulate only during the growing season and after the decisions of what and how much to plant, this relationship is considered spurious and likely traceable to the joint influences of these variables on how much grain is produced in a given year.

In summary, one must be cautious in assigning dependency relationships among input variables of a system based on bivariate correlation analysis. To be safe, only those dependencies that are intuitively logical should be taken into account. Accordingly, for the Monte Carlo simulation of the next section only the statistically significant dependencies highlighted in bold in Table VI-13 are incorporated.

MONTE CARLO SIMULATION OF TOP-DOWN SYSTEM

The probability distributions and/or values of variables in the top-down analysis as defined above were coded into an @Risk© spreadsheet for the Monte Carlo simulation. The particular version of @Risk© used is an add-in application to the EXCEL spreadsheet program. Setting up the simulation entailed defining and arranging the inputs (i.e., the probability distributions for the top-down variables), identifying correlations among the inputs (i.e., the important correlations described in the preceding paragraphs), and defining outputs (i.e., the calculation of annual grain production and number of tows from the sampled probability distributions of the input parameters). The program was also told to monitor convergence and automatically stop when the simulation statistics had converged.[38]

Table VI-14 shows an example of the simple structure of the @Risk© spreadsheet for the simulation of the forecast for the year 2005. By inspecting the table, one will notice that both the structure of the master equation (6.4) and the assumed probability distributions are built into the spreadsheet. As shown, the probability distributions are referenced by name or abbreviation, and are defined by certain parameters that define their respective shapes (the parameters shown have been rounded to economize on space). For example, the triangular distribution for the total number of acres planted is referenced as "RiskTriang" and is defined by the three parameters, which correspond to the minimum, most likely, and maximum values that are assigned, respectively (see Table VI-1). Notice that Cell B9 of the spreadsheet contains the most likely value. Similarly, the probability distributions for all other inputs are referenced by name with

[38] It is left up to the reader to learn further details about the @Risk© program and the range of functions and procedures it provides for risk and uncertainty analysis.

TABLE VI-14

EXAMPLE OF @Risk© SIMULATION SPREADSHEET

	A	B	C	D	E	F	G	H	I	J
1	Tot Acres	cdd	corn yield regression equation				wheat yield regression equation			
2	RiskTriang(16600000,B9,22000000)	RiskNormal(1572.92,158.36)	c-int	c-time	c-cdd	c-ran. error	w-int	w-time	w-cdd	w-ran. error
3			RiskNormal(78.09,34.31)	RiskNormal(1.01,0.44)	RiskNormal(0.029,0.02)	RiskNormal(0,16.7)	RiskNormal(29.6,11.6)	RiskNormal(0.31,0.15)	RiskNormal(0.01,0.01)	RiskNormal (0,5.6)
4			soy yield regression equation				other yield regression equation			
5			s-int	s-time	s-cdd	s-ran. error	o-int	o-time	o-cdd	o-ran. error
6			RiskNormal(29.34,7.07)	RiskNormal(0.4,0.09)	RiskNormal(0.01,0.004)	RiskNormal(0,3.4)	RiskNormal(34.9,13.9)	RiskNormal(0.36,0.18)	RiskNormal(0.02,0.01)	RiskNormal(0,6.7)
7	year	2005								
8	year-1995	10								
9	acres expected	20500000								
10								tons by Chad. Lock	barges by Lock	
11	crop/year	ptotac	ptotacadj	yield	production	pexport	pbarge			
12	corn	RiskWeibull(36.2,0.52)	B12/$B$16	(C3+(B8*D3)+($B$2*E3)+F3)	($A$2*C12)*D12	RiskNormal(0.6,0.05)	RiskUniform(0.6,0.8)	(E12*F12*G12)*56/2000	H12/1520	
13	soy	RiskNormal(0.358,0.0357)	B13/$B$16	(C6+(B8*D6)+($B$2*E6)+F6)	($A$2*C13)*D13	RiskWeibull(11.0,0.4)	RiskUniform(0.6,0.7)	(E13*F13*G13)*60/2000	H13/1520	
14	wheat	RiskNormal(0.0675,0.0127)	B14/$B$16	(G3+(B8*H3)+($B$2*I3)+J3)	($A$2*C14)*D14	RiskNormal(0.7,0.06)	RiskUniform(0.8,0.9)	(E14*F14*G14)*60/2000	H14/1520	
15	other	RiskLognorm(0.0621,0.0277)	B15/$B$16	(G6+(B8*H6)+($B$2*I6)+J6)	($A$2*C15)*D15	RiskWeibull(6.2,0.2)	RiskUniform(0.4,0.6)	(E15*F15*G15)*60/2000	H15/1520	
16	sums	SUM(B12:B15)								
17									SUM(I12:I15)	total barges
18									I17/10	total tows

the necessary parameters that define their shape. The cells for yield, production, number of barges, and number of tows are calculated fields (or outputs). The cell for total tows at the bottom right-hand corner of the table represents the simulated output of interest. For every iteration of a simulation, the probability distributions are sampled randomly to re-calculate the number of tows. @Risk© stores these values so that one may define a probability distribution on tows. Further, within the Monte Carlo process, the sampling from one probability distribution is allowed to affect the sampling from one or more other probability distributions, based on the specification of dependencies among the variables. These dependencies may be entered directly into @Risk© using the software's *Correlate* option. Table VI-15 corresponds exactly to the correlation matrix specified within @Risk© for the simulation.[39]

Separate samplings of 4,000, 2,700, and 3,400 iterations were performed for the forecast years 2005, 2010, and 2015, respectively, after which the simulations of the top-down system converged. The resulting distributions on tows are reported in Table VI-16 in the form of percentiles. The three panels of Figure VI-7 graph the simulated probability distributions for future number of tows for each of the forecast years. The distributions are bell-shaped, but not strictly normal.

The expected values reported in the table reflect point predictions for the future number of tows. As shown, the expected annual number of tows is forecast to rise from 1,341 in 2005 to 1,453 in 2015. Referring back to the assumptions that were made for the analysis, this reflects the anticipated growth in acres planted and improvements in grain yield over time.

One can derive confidence intervals (i.e., the uncertainty) on the forecast number of tows using the percentiles reported in Table VI-16. A 90 percent interval would be formed by the 5th and 95th percentile values, since together only 10 percent of all cases would be expected to fall below or above these values, respectively. Table VI-17 summarizes the 90 percent prediction intervals for the top-down forecast with uncertainty.

SUMMARY

This chapter explained the top-down approach to forecasting as one that attempts to define all of the factors that influence the forecast variable of interest. In this regard, the top-down methodology is really a process of thought that uncovers many possible sources of uncertainty in the forecasting process. Within the example provided for grain production and river shipments in the Evanstown/Oak River region, it was discovered that the real underlying mechanisms that determine grain production and the number of tows represent a complex mix of physical and socioeconomic factors that are much easier to describe than to estimate. Even in a detailed top-down analysis, simplifications must be made because of lack of knowledge, data, time, or budget. Within these constraints, the top-down forecast with uncertainty can be characterized as a step-

[39]Notice the correlations among the coefficients of the yield regression models. These account for the covariance among the parameter estimates.

TABLE VI-15

CORRELATION MATRIX ENTERED WITHIN @RISK©

	ACTOT	CDD	c-yield-int	c-yield-time	c-yield-cdd	c-yield-r. err.	w-yield-int	w-yield-time	w-yield-cdd	w-yield-r. err.	s-yield-int	s-yield-time	s-yield-cdd	s-yield-r. err.	p-yield-int	p-yield-time	p-yield-cdd	p-yield-r. err.	FACORN	FECORN	FBCORN	FASOY	FESOY	FBSOY	FAWHET	FEWHET	FBWHET	FAOTH	FEOTH	FBOTH
ACTOT	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
CDD	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c-yield-int	0	0	1	0.2508	-0.9828	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c-yield-time	0	0	0.2508	1	-0.0946	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c-yield-cdd	0	0	-0.9828	-0.0946	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
c-yield-r. err.	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
w-yield-int	0	0	0	0	0	0	1	0.2508	-0.9828	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
w-yield-time	0	0	0	0	0	0	0.2508	1	-0.0946	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
w-yield-cdd	0	0	0	0	0	0	-0.9828	-0.0946	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
w-yield-r. err.	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s-yield-int	0	0	0	0	0	0	0	0	0	0	1	0.2508	-0.9828	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s-yield-time	0	0	0	0	0	0	0	0	0	0	0.2508	1	-0.0946	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s-yield-cdd	0	0	0	0	0	0	0	0	0	0	-0.9828	-0.0946	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
s-yield-r. err.	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
p-yield-int	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.2508	-0.9828	0	0	0	0	0	0	0	0	0	0	0	0	
p-yield-time	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2508	1	-0.0946	0	0	0	0	0	0	0	0	0	0	0	0	
p-yield-cdd	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.9828	-0.0946	1	0	0	0	0	0	0	0	0	0	0	0	0	
p-yield-r. err.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
FACORN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	-0.6602	0	0	0	0	0	0	0	
FECORN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.4253	0	0	0.9283	0	0	0.4185	
FBCORN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.4422	0	0	
FASOY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.6602	0	0	1	0	0.4232	0	0	0	-0.7294	0	
FESOY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4253	0	0	1	0.47397	0	0	0	0	0.38352	
FBSOY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.42322	0.47397	1	0	0	0.59839	0	0	
FAWHET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
FEWHET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.92826	0	0	0.48676	0	0	1	0	0	0.34646	
FBWHET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.44218	0	0	0.59839	0	0	1	0	0	
FAOTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.7294	0	0	0	0	0	1	-0.5447	-0.4079
FEOTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.41848	0	0	0.38352	0	0	0.34646	0	-0.5447	1	

TABLE VI-16
SIMULATION RESULTS: FORECAST NUMBER OF TOWS

	Forecast Year		
Statistic	**2005**	**2010**	**2015**
Expected Value	1,341	1,398	1,453
Standard Deviation	287	293	306
Minimum	585	547	575
Maximum	2,718	2,554	2,818
Iterations	4,000	2,700	3,400
Percentiles			
5%	909	941	981
10%	988	1,033	1,076
15%	1,041	1,092	1,145
20%	1,097	1,149	1,195
25%	1,140	1,190	1,242
30%	1,177	1,236	1,289
35%	1,215	1,273	1,327
40%	1,251	1,312	1,360
45%	1,285	1,351	1,394
50%	1,316	1,387	1,431
55%	1,350	1,420	1,469
60%	1,393	1,452	1,510
65%	1,434	1,490	1,555
70%	1,475	1,534	1,594
75%	1,524	1,577	1,636
80%	1,572	1,641	1,693
85%	1,630	1,707	1,761
90%	1,728	1,783	1,848
95%	1,853	1,899	1,995

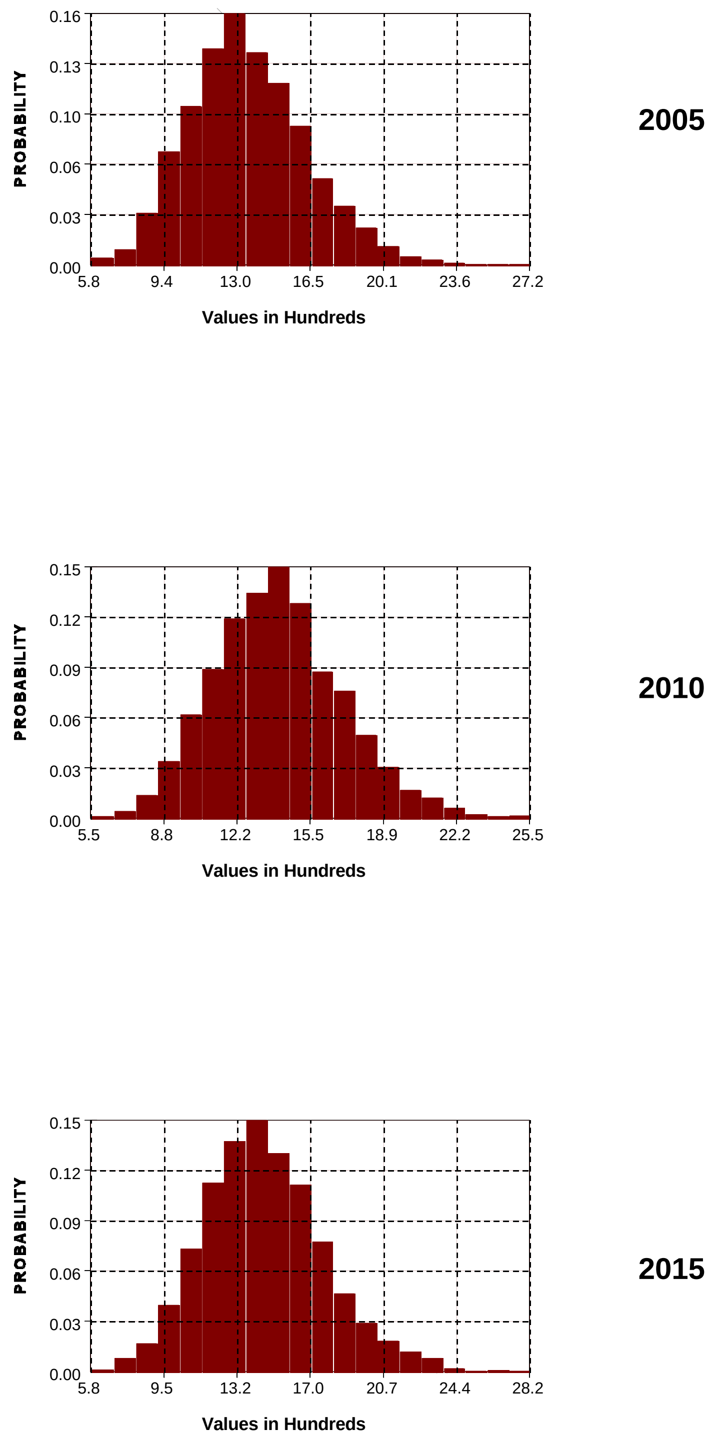

FIGURE VI-7

**SIMULATED FORECAST DISTRIBUTION
FOR ANNUAL NUMBER OF TOWS**

TABLE VI-17

THE TOP-DOWN FORECAST WITH UNCERTAINTY: 90 PERCENT CONFIDENCE INTERVAL OF FORECAST OF TOWS PASSING CHADWICK LOCK

	Forecast Scenario		
Forecast Year	**Lower Bound**	**Expected**	**Upper Bound**
2005	909	1,341	1,853
2010	941	1,398	1,899
2015	981	1,453	1,995

wise process of determining a system of inputs. How these variable inputs are distributed and interact with one another determines the distribution of possible outcomes.

The Monte Carlo simulation method was shown to provide an efficient means of constructing a probability distribution of forecast tows. Application of the method was discussed to require the development of a functional relation between the number of tows and a set of causal parameters, as well as assumptions regarding the probability distributions and interdependence among the causal variables. Where possible, the distributions of important parameters were defined according to theory or functional relationships. In cases where knowledge of underlying distributions were lacking, assignments was made with the aid of the BestFit© probability distribution fitting software together with professional judgement. Finally, @Risk© simulation software was relied upon to implement the simulation procedure.

VII. CHOOSING A FORECASTING METHODOLOGY

The last four chapters of this manual presented ways to estimate and portray risk and uncertainty in forecasts of waterborne commodity movements within the context of four distinct forecasting methodologies. The forecasting methodologies differ considerably, and so too do the prescribed ways of incorporating uncertainty analysis. It seems fitting, then, to conclude this manual with some discussion on choosing the appropriate forecasting methodology.[40]

CONSTRAINTS, ACCURACY, AND TRADEOFFS

The choice of forecasting methodology is clearly constrained by limits on time and budget, or, in other words, by the resources that are available for collecting data, estimating models, performing uncertainty analysis, etc. For example, if one cannot afford to collect data for or hire someone to perform a detailed top-down analysis, then the top-down approach is not a choice. On the other hand, if one could afford to choose any one of the four forecasting methodologies presented in this manual (i.e., if all methodologies fall in the realm of choice), which methodology should he or she choose? The answer to this question is less obvious than one might think. A customary answer would be to select the methodology that produces the most accurate (or least wrong) forecast. A forecasting methodology that produces accurate forecasts would obviously be preferred to one that does not.[41] However, accuracy may determined only after a forecasting approach has been selected and a forecast prepared. Unfortunately, one typically does not have the luxury nor the time to test the performance of different methodologies in order to choose the one that is most accurate. Even if one were afforded this luxury, there is no guarantee that the selected methodology would remain more accurate under different conditions.

There is a common belief that there is a tradeoff between the cost of a forecasting methodology and its accuracy. Two distinct notions tend to underlie this belief. First, more sophisticated and complex forecasting methodologies are generally considered to be more costly to implement. Second, more sophisticated methodologies are generally assumed to produce more accurate forecasts. The first proposition has some merit. For example, the cost of a top-down analysis would probably exceed the costs of a growth-rate or trend analysis. The second proposition, however, is much more speculative, especially if one considers the track record of forecasters. Evidence from the field of economic forecasting suggests that forecasts tend to be inaccurate, regardless of methodology. One survey of the literature on this subject has even suggested that

[40] Keep in mind that this report has concerned the use of only four of many methodologies that can be conceived for forecasting commodity flows.

[41] The accuracy of a forecasting methodology should be differentiated from the accuracy of a particular forecast. Unlike the former, the latter is measurable numerically by some standard formulae (e.g., see Kennedy, 1992)

simple methods are as accurate as sophisticated methods (Mahmoud, 1984). Thus, the jury is still out on the cost versus accuracy issue. Forecasting accuracy is an evasive issue that cannot in most cases determine the choice of methodology.

Certainly, this does not mean that forecasting accuracy is unimportant. If one is indeed able to conclude that one approach to forecasting is more accurate than another, then, by all means, it should be selected. Neither does this imply that it is acceptable, for example, to base the decision to fund a billion-dollar waterway improvement on the findings of a traffic forecast that is based on historical growth rates. In fact, it is unlikely that anyone *would* base such a weighty decision on this technique, *even if* application of the technique was shown to produce a narrow confidence interval on the forecasted variable.[42] Why would this approach be insufficient to support such a decision? Because, this approach does not define for decisionmakers any of the factors that determine waterway traffic in the real world. Would someone use this approach as a basis for planning manpower staffing at a lock over the next five years? The answer probably depends on whether the shortcomings of the approach and the uncertainties in the forecast have been defined and understood by the appropriate planning authority.

Depending on the situation, similar stories could be told of all of the other forecasting techniques reviewed herein. To some, the shippers' survey of Chapter IV may require too much faith in subjective judgement to be of much use to long-term planning. In the same vein, some decisionmakers might not be willing to place much confidence in a forecast derived from a multiple regression model like the one of Chapter V, since it does not (or cannot) incorporate all of the variables that influence waterborne commerce. One may even consider the assumptions of a top-down analysis like the one of Chapter VI to be too simplistic and unrepresentative of reality.

The point of this discussion is that there is no preconceived answer to the leading question of which forecasting methodology to select. Ultimately, the choice of methodology must be dictated by circumstance and is a function of (1) the decision-support objectives of the forecast, (2) budget and time constraints, and (3) *who* determines the constraints. As such, this manual does not endorse any particular methodology per se. However this manual does very strongly endorse the thought process that is inherent in the top-down approach and to a lesser extent in the multiple regression approach.[43] These methodologies attempt to describe the workings of the system that determines values of the variable of interest, such as grain shipment down the Oak River or the number of tows passing Chadwick Lock. The representation of this thought process is meaningful and appealing to decisionmakers, because it alerts them of what is known, what is unknown or assumed, and what is uncertain.

---

[42] Keep in mind also that narrow forecast confidence intervals do not necessarily imply a good methodology. Neither do wide confidence bands necessarily imply a bad methodology. A good methodology would be considered as one that attempts to identify and account for as many sources of uncertainty as possible. As a result, the prediction interval derived from the "good" methodology may be considerably wider than the interval of a "bad" methodology that did not attempt to uncover so many sources of uncertainty.

[43] Note that the top-down approach can be made up of a system of regression equations.

SOME BASIC RULES

The purpose of this manual was to define ways in which one might measure and portray forecast uncertainty within the context of four standard Corps forecasting methodologies. As was shown, the portrayal of forecast uncertainty truly requires one to understand what drives the forecasted variable. The thought process of defining causality may be more informative than the actual forecast itself.

If one considers that the purpose of forecasting is to support decisions regarding the allocation of resources, then the role of the forecaster should be to make sure that decisionmakers understand the complexity of the real world and the uncertainties of the forecasting process. Following four basic rules will help one fulfill this role, regardless of choice of methodology:

Rule 1. *Explain the process by which the real-world outcomes are produced.* This requires one first to understand the process he or she is forecasting, and is essential for determining the sources of uncertainty in the forecast. A forecasting exercise is incomplete without an explicit discussion of causal factors.

Rule 2. *Identify the potential sources of error and attempt to estimate and portray the degree of error.* This requires a full understanding of the process that produces the data you have to work with and some tools and techniques for quantifying and representing uncertainty. This manual may be used to help one define techniques to quantify uncertainty. The techniques presented in this manual are only some of many possible ways of measuring and portraying uncertainties in the techniques described.

Rule 3. *Rely on theory in the absence of or lack of data*. This is not as trivial as it might sound. Sometimes one does not readily have data to portray and measure uncertainty in a variable. Economic and scientific theory may shed light on the real world data generating process and therefore may help one form appropriate assumptions. As shown in this manual, statistical and mathematical theory may be relied upon to facilitate the measurement and treatment of uncertainty, even if one cannot pinpoint a theory that supports evidence provided by data.

Rule 4. *Make all assumptions explicit.* Even in the most well-funded and long-lasting study, assumptions have to be made. As suggested in this manual, more assumptions may be a result of a better understanding of the workings of the system. By <u>listing</u> assumptions, one will enlighten the forecasting audience of what can be done for a given amount of money and an allotted amount of time.

These rules are intended to maximize both the *amount* of information derived from a forecasting exercise and the *flow* of information to those who are interested in the results. Finally, and as mentioned earlier in the manual, it is important to point out the possible shortcomings of a forecast or forecasting model, even if it seems adequate, and particularly, if time and budget constraints limit its refinement.

REFERENCES

Ayyub, B.M., B. Riley, and M. Hoge. 1996. *Expert Elicitation of Unsatisfactory-Performance Probabilities and Consequences for Civil Works Facilities*. Washington, DC. Prepared for Headquarters, U.S. Army Corps of Engineers.

Beim, G.K. and B.F. Hobbs. 1994. *Poe Lock System Risk Analysis*. Fort Belvoir, VA. Prepared for The Institute for Water Resources, U.S. Army Corps of Engineers.

Grier, D. W. and L. L. Skaggs. 1992. *A Review of 16 Planning and Forecast Methodologies Used in U.S. Army corps of Engineers Inland Navigation Studies*. IWR Report 92-R-4. Ft. Belvoir, Virginia, U.S. Army Corps of Engineers, Institute for Water Resources.

Guide to Using BestFit© Probability Distribution Fitting Software for Windows©. February 1996. New Field, New York: Palisade Corporation. (For more information about @Risk© one may (1) telephone, (607) 277-8000; (2) fax (607) 277-8001, or (3) e-mail, *info@palisade.com*)

Guide to Using @Risk©: Risk Analysis and Simulation Add-In for Microsoft© Excel or Lotus© 1-2-3. Windows© Version. March 1996. New Field, New York: Palisade Corporation. (For more information about @Risk© one may (1) telephone, (607) 277-8000; (2) fax (607) 277-8001, or (3) e-mail, *info@palisade.com*)

Harnett, D. L. and J. L. Murphy. 1985. *Statistical Analysis for Business and Economics*. Third Edition. Reading, Massachusetts: Addison-Wesley Publishing Company.

Harrington, K. W. and J. S. Willett. 1996. *Transport Mode Selection and Inland Waterborne Commerce: An Annotated Bibliography*. Carbondale, IL: Planning and Management Consultants, Ltd.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee. 1988. *Introduction to the Theory and Practice of Econometrics*. Second Edition. New York: John Wiley.

Kennedy, P.E. 1992. *A Guide to Econometrics*. Third Edition. Cambridge, Massachusetts: The MIT Press.

Kexel, D.T. 1988. “The Ten Year Forecast - How Certain Are We?” Paper presented at EPRI’s Eight Electric Utility Forecasting Symposium, October 23-25, 1991, Baltimore, Maryland.

Kmenta. 1986. *Elements of Econometrics*. Second Edition. New York: Macmillian Publishing Company.

Kocik, J. W., and P. Feinsilver. 1993. “On Error Variance for Forecast Models with Uncertain Explanatory Variables.” Preprint of Mathematics Department. Carbondale, IL: Southern Illinois University - Carbondale.

Kocik, J. W., J. C. Kiefer, and B. Dziegielewski. 1993. *Uncertainty in the MWD-Main Water Use Forecast: The Development and Application of a Methodology for Forecasting Error Estimation*. Carbondale, IL: Planning and Management Consultants, Ltd.

Mahmoud, E. 1984. "Accuracy in Forecasting: A Survey." *Journal of Forecasting*. 3:139-59.

Pindyck, R.S. and D. L. Rubinfeld. 1981. *Econometric Models & Economic Forecasts*. Second Edition. McGraw-Hill.

SAS©/STAT© User's Guide, Version 6. Second Edition. 1993. Cary, North Carolina: SAS Institute Inc. (For more information about SAS© one may telephone (919) 677-8000).

U.S. Water Resources Council. 1983. *Economic and Environmental Principles and Guidelines for Water and Related Land Resources Implementation Studies*. Washington, D.C.: U.S. Water Resources Council.

APPENDIX A

HISTORICAL DATA FOR TOP-DOWN INPUT VARIABLES

TABLE A-1
HISTORICAL DATA FOR TOP-DOWN INPUT VARIABLES

	Acres Planted					Fraction of planted acres				Yield (bushels) per acre			
YEAR	corn	wheat	soybean	other	total	corn	wheat	soybean	other	corn	wheat	soybean	other
1970	8,763,571	639,667	5,170,000	2,093,000	16,666,238	0.5258	0.0384	0.3102	0.1256	110.0	32.3	30.6	50.6
1971	8,544,286	870,000	5,143,333	2,033,333	16,590,952	0.5150	0.0524	0.3100	0.1226	113.7	41.0	29.2	58.0
1972	8,830,000	910,333	5,581,667	1,662,000	16,984,000	0.5199	0.0536	0.3286	0.0979	109.2	38.1	33.7	52.9
1973	9,403,571	1,114,667	7,023,333	1,746,667	19,288,238	0.4875	0.0578	0.3641	0.0906	102.6	35.3	32.4	52.9
1974	9,964,286	1,528,667	6,470,000	1,566,667	19,529,619	0.5102	0.0783	0.3313	0.0802	76.9	29.4	26.1	50.0
1975	10,375,000	1,565,667	6,313,333	1,578,333	19,832,333	0.5231	0.0789	0.3183	0.0796	95.7	33.6	33.6	50.7
1976	10,746,429	2,012,000	5,676,667	1,592,000	20,027,095	0.5366	0.1005	0.2834	0.0795	91.2	34.5	30.2	50.0
1977	10,707,143	1,668,667	6,566,667	1,736,333	20,678,810	0.5178	0.0807	0.3176	0.0840	95.9	40.8	36.6	62.2
1978	10,789,286	1,233,667	6,950,000	1,422,667	20,395,619	0.5290	0.0605	0.3408	0.0698	111.3	34.3	35.4	54.4
1979	10,789,286	1,296,000	7,656,667	1,196,667	20,938,619	0.5153	0.0619	0.3657	0.0572	121.6	37.6	36.9	58.2
1980	11,082,143	1,610,333	7,460,000	1,192,667	21,345,143	0.5192	0.0754	0.3495	0.0559	101.1	37.9	34.9	55.3
1981	11,407,143	1,861,667	7,183,333	1,248,333	21,700,476	0.5257	0.0858	0.3310	0.0575	122.2	43.5	37.5	61.1
1982	11,103,571	1,594,667	7,476,667	1,223,667	21,398,571	0.5189	0.0745	0.3494	0.0572	122.6	41.4	37.0	60.7
1983	7,435,714	1,196,667	7,203,333	1,082,333	16,918,048	0.4395	0.0707	0.4258	0.0640	83.3	40.4	32.3	54.5
1984	10,814,286	1,417,667	7,553,333	1,113,333	20,898,619	0.5175	0.0678	0.3614	0.0533	111.7	45.8	31.9	65.3
1985	11,150,000	1,181,667	7,383,333	1,188,333	20,903,333	0.5334	0.0565	0.3532	0.0568	127.1	51.9	38.5	71.5
1986	10,017,857	1,231,333	7,433,333	960,000	19,642,524	0.5100	0.0627	0.3784	0.0489	132.4	38.4	39.2	59.0
1987	8,625,000	1,166,333	7,083,333	883,333	17,758,000	0.4857	0.0657	0.3989	0.0497	130.2	45.7	40.3	59.2
1988	8,928,571	1,178,333	7,200,000	786,667	18,093,571	0.4935	0.0651	0.3979	0.0435	77.9	34.2	28.3	39.0
1989	10,214,286	1,516,333	7,376,667	913,333	20,020,619	0.5102	0.0757	0.3685	0.0456	121.3	46.4	39.0	63.1
1990	10,339,286	1,596,667	7,200,000	831,667	19,967,619	0.5178	0.0800	0.3606	0.0417	125.9	48.2	39.9	66.5
1991	10,428,571	1,201,667	7,693,333	721,000	20,044,571	0.5203	0.0599	0.3838	0.0360	113.8	31.6	38.4	49.9
1992	10,892,857	1,331,667	7,666,667	646,667	20,537,857	0.5304	0.0648	0.3733	0.0315	140.7	51.1	40.7	75.3
1993	9,142,857	1,296,000	7,433,333	550,000	18,422,190	0.4963	0.0703	0.4035	0.0299	97.7	36.0	34.0	55.8
1994	10,928,571	1,172,333	7,966,667	583,333	20,650,905	0.5292	0.0568	0.3858	0.0282	151.4	35.3	46.7	60.0
1995	9,821,429	1,209,333	8,233,333	503,333	19,767,429	0.4968	0.0612	0.4165	0.0255	117.6	44.7	41.3	63.6

TABLE A-1 (Continued)
HISTORICAL DATA FOR TOP-DOWN INPUT VARIABLES

YEAR	Fraction of production exported: corn	wheat	soybean	other	Fraction of exports shipped by barge: corn	wheat	soybean	other	Number of tows	Relative price	CDD
1970	0.5400	0.6200	0.3470	0.1350	0.7700	0.9100	0.6900	0.3700	802	0.760	1425
1971	0.5500	0.6600	0.3740	0.2350	0.8300	0.9300	0.7000	0.4300	887	0.819	1447
1972	0.4800	0.5560	0.2250	0.1200	0.7600	0.8860	0.6600	0.3900	704	0.838	1500
1973	0.5800	0.6830	0.3050	0.2050	0.7300	0.8400	0.6300	0.4500	817	0.895	1449
1974	0.6000	0.7100	0.3810	0.2370	0.8000	0.9110	0.6750	0.5000	736	0.769	1330
1975	0.6500	0.7900	0.3220	0.1630	0.7560	0.8540	0.6500	0.4600	975	0.650	1625
1976	0.5500	0.6450	0.2880	0.1910	0.7420	0.8490	0.6600	0.4300	800	0.820	1725
1977	0.5800	0.7000	0.3470	0.1450	0.7400	0.8330	0.6200	0.3700	882	0.802	1800
1978	0.5900	0.7200	0.3930	0.1480	0.8200	0.8770	0.7100	0.5100	1,162	0.697	1825
1979	0.6300	0.7440	0.3760	0.2610	0.7780	0.8520	0.6870	0.3500	1,286	0.559	1550
1980	0.6400	0.6870	0.3340	0.2390	0.7820	0.8550	0.6900	0.5300	1,121	0.751	1445
1981	0.6600	0.7920	0.3210	0.1580	0.7690	0.8420	0.6650	0.4500	1,415	0.851	1675
1982	0.6200	0.7390	0.3150	0.2390	0.7270	0.7870	0.6400	0.3900	1,227	0.703	1681
1983	0.5800	0.6960	0.3300	0.2240	0.7320	0.8250	0.6700	0.5400	526	0.901	1375
1984	0.5500	0.6650	0.3290	0.2260	0.7550	0.8660	0.7140	0.5000	1,003	0.611	1895
1985	0.6800	0.8160	0.4070	0.2250	0.7560	0.8770	0.7200	0.4400	1,457	0.865	1675
1986	0.6400	0.7680	0.3840	0.2720	0.6570	0.8000	0.6880	0.3900	1,115	0.896	1700
1987	0.5900	0.6600	0.3590	0.2560	0.6330	0.8140	0.6890	0.5700	839	0.805	1695
1988	0.5800	0.7400	0.3750	0.1900	0.6990	0.8260	0.7000	0.5500	564	0.802	1320
1989	0.6500	0.7800	0.3890	0.2850	0.7180	0.9210	0.7320	0.4500	1,156	0.964	1450
1990	0.6400	0.7520	0.3230	0.2600	0.6980	0.9000	0.7200	0.4200	1,163	1.079	1395
1991	0.6300	0.7560	0.3720	0.2200	0.6750	0.9000	0.7130	0.4400	1,010	0.952	1449
1992	0.6200	0.7580	0.3120	0.2500	0.7000	0.9390	0.7250	0.4500	1,330	0.997	1550
1993	0.5700	0.6840	0.3370	0.2420	0.8310	0.9500	0.7270	0.4700	847	0.962	1650
1994	0.5500	0.6600	0.3290	0.2260	0.7610	0.9250	0.6900	0.6000	1,385	0.970	1580
1995	0.5400	0.6610	0.3860	0.2200	0.7470	0.8880	0.7190	0.4900	932	0.950	1685